Dynamics of macromolecular complexes through computational modelling and structural mass spectrometry

Lau, Andy Man Chung

Awarding institution:
King's College London

King's College London

Dynamics of macromolecular complexes through computational modelling and structural mass spectrometry

Andy M. C. Lau

This thesis is submitted for the degree of

Doctor of Philosophy

to

King's College London

University of London

School of Natural and Mathematical Sciences

Department of Chemistry

2019

"It is structure that we look for whenever we try to understand anything. All science is built upon this search; we investigate how the cell is built of reticular material, cytoplasm, chromosomes; how crystals aggregate; how atoms are fastened together; how electrons constitute a chemical bond between atoms. We like to understand, and to explain, observed facts in terms of structure."

- Linus Pauling

"Everything that living things do can be understood in terms of the jigglings and wigglings of atoms."

- Richard Feynman

ABSTRACT

For a large part its history, structural biology has relied on X-ray crystallography as the primary means of accessing the atomic structures of proteins and other molecules. 140,000 crystal structures currently populate the Protein Data Bank, each representing a single snapshot of the cellular landscape. To what extent are these conformations representative of each protein's biochemical capabilities? What percentage of the proteome's conformational space has been mapped? While these questions remain far too large and impossible for any one PhD student to answer, this thesis demonstrates several ways in which the conformational dynamics of proteins and complexes can be tackled through combining computational modelling with the powerful analytical capabilities of mass spectrometry (MS).

The structural characterisation of large dynamic molecules remains one of the major challenges in biology due to the lack of techniques capable of capturing their motions. In Chapter 2 of this thesis, we model the conformational dynamics of large flexible Immunoglobulin G (IgG) antibodies using a workflow integrating ion mobility (IM)-MS, modelling and molecular dynamics simulations. This workflow provides the means of leveraging experimental IM-MS measurements with atomistic models. Next, the use of hydrogen deuterium exchange (HDX)-MS for protein characterisation has gained traction over the recent years due to its ability to capture dynamic aspects of protein behaviour. Advances in HDX-capable instrumentation has been paralleled by developments into software that facilitates its analysis. Chapter 3 documents the development of Deuteros, a software designed for rapid statistical analysis and visualisation of HDX-MS data, including recent upgrades to its code and analytical facilities. Finally, in Chapter 4, the structural dynamics of a large multi-subunit enzyme known as the Constitutive Photomorphogenesis 9 Signalosome (CSN) is examined using the synergy of structural MS and cryo-electron

microscopy density maps. Integration of these techniques provided a method of exposing the multifaceted dimensions of the CSN and revealed its stepwise activation cascade. The work comprising Chapters 2, 3 and 4 have each been published in *Angewandte Chemie*, *Bioinformatics* and *Nature Communications* respectively. Overall, the research presented in this thesis has not only contributed important biological insights into the conformational dynamics of IgG and the CSN but has also demonstrated the utility of using integrative approaches for the characterisation of dynamic macromolecules.

ACKNOWLEDGEMENTS

Over the past three and half years, I have had the pleasure of working with many talented people who have had a significant impact on my development both as a scientist and as an individual. While protein-protein interactions can be captured using various structural techniques, human-human interactions can only be measured through emotional bonds and not hydrogen.

First and foremost, I owe my thanks to my supervisor, Dr Argyris Politis, who has been a guiding presence throughout my PhD. In particular, I want to express my gratitude to my colleagues and friends Zainab Ahdash and Dr Chloe Martens who have made my experience at King's so enjoyable, both professionally and socially. I want to thank KJ Hansen whom I worked alongside for the IgG modelling project, for always being able to answer my mass spectrometry questions and for all the late-night conversations about IgG, Deuteros and HDX. My thanks to Matthew Harris for the days spent designing publication covers, both successful and unsuccessful, and Euan Pyle for making my PhD so much more entertaining. I would also like to thank Dr Carla Schmidt, Dr Sarah Faull and Dr Ed Morris who I had the pleasure of working alongside for the CSN project. Without them, a large portion of my PhD would not have been possible. Special thanks to my past supervisors, Dr Lindsay McDermott, Henna Zahid and Prof. Stephen Perkins at UCL, for introducing me to the world of structural biology and who's guidance has led to an extremely rewarding experience. Last but not least, I would like to acknowledge the BBSRC London Interdisciplinary Doctoral Programme (LIDo) for converting me from an inexperienced bachelor student to a less inexperienced PhD student, and thank Nadine Mogford and the rest of the LIDo team at UCL, who have been tremendous in keeping me on track throughout the years.

Finally, I am grateful to my partner Wei-Hong Tseng for the exceptional care, emotional support, and for putting up with me over the last few months of writing this thesis, my close friend Anka Lucic for the journey through science and education together, and lastly my parents for their overwhelming support and encouragement.

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

ABBREVIATIONS

MeCN	Acetonitrile
ADH	Alcohol dehydrogenase
AGC	Automatic gain control
AN	Accession number
ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
APC	Anaphase promoting complex
ATP	Adenosine triphosphate
BC	Elongin B-Elongin C
BPTI	Bovine pancreatic trypsin inhibitor
BTB	Broad-complex, Tramtrack, Bric-a-brac
CCS	Collisional cross section
CHARMM	Chemistry at Harvard molecular mechanics
CI	Confidence interval
CID	Collision induced dissociation
COP9	Constitutive photomorphogenesis 9
CRISPR	Clustered regularly interspaced short palindromic repeats
CRL	Cullin ring ligase
CRM	Charged residue model
CSN	Constitutive photomorphogenesis 9 signalosome
CTD	C-terminal domain
CUL	Cullin
CYS	Cysteine
DCAF	Ddb1-cul4a-associated factor
DNA	Deoxyribonucleic acid
DOPE	Discrete optimised protein energy
DT	Drift tube
DTT	Dithiothreitol
DU	Deuterium uptake
EDTA	Ethylenediaminetetraacetic acid
EHSS	Exact hard sphere scattering
ELOB	Elongin B
ELOC	Elongin C
EM	Electron microscopy
EMDB	Electron microscopy data bank
ESI	Electrospray ionisation
ETD	Electron transfer dissociation
FA	Formic acid

FSC	Fourier shell correlation
GDH	Glutamate dehydrogenase
GROMACS	Groningen machine for Chemical Simulations
GUI	Graphical user interface
HCD	Higher energy collisional dissociation
HDMS	High definition mass spectrometry
HDX	Hydrogen-deuterium exchange
HECT	Homologous to the E6-AP Carboxyl Terminus
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HIF	Hypoxia inducible factor
HIV	Human immunodeficiency virus
HPLC	High-performance liquid chromatography
HRE	Hypoxia-response element
IL	Interleukin
IM	Ion mobility
IMP	Integrative modelling platform
IMPACT	Ion mobility projection approximation calculation tool
IPTG	Isopropyl-β-D-thiogalactoside
IQR	Interquartile range
JAMM	JAB1/MPN/MOV34
LC	Liquid-chromatography
LDS	Lithium dodecyl sulphate
LINCS	Linear constraint solver
LRR	Leucine-rich-repeat
MALDI	Matrix assisted laser desorption/ionization
MD	Molecular dynamics
MDFF	Molecular dynamics flexible fitting
MGC	Mammalian gene collection
MPN	Mpr1, Pad1 N-terminal
MS	Mass spectrometry
MSE	Mean squared error
MW	Molecular weight
MWCO	Molecular weight cut off
NAE	NEDD8 activating enzyme
NAMD	Nanoscale molecular dynamics
NEDD8	Neural precursor cell expressed developmentally down-regulated protein 8
NF-κB	Nuclear Factor kappa-light-chain-enhancer of activated B cells
NHS	N-hydroxysuccinimide
NMR	Nuclear magnetic resonance
OPLS	Optimized Potential for Liquid Simulations

PA	Projection approximation
PARC	p53 cytoplasmic anchor
PDB	Protein data bank
PLGS	Protein lynx global server
PLIMSTEX	Protein–ligand interactions by mass spectrometry, titration, and H/D exchange
PME	Particle mesh Ewald
POTRA	Polypeptide-transport-associated
PRAME	Preferentially expressed antigen of melanoma
PROTAC	Proteolysis targeting chimera
PSA	Projected superposition approximation
QTOF	Quadrupole-Time-of-flight
RBX	RING box
RFU	Relative fractional uptake
RING	Really interesting new gene
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
RNA	Ribonucleic acid
RRT	Rapidly-exploring random tree
RT	Retention time
SASA	Solvent accessible surface area
SAXS	Small angle x-ray scattering
SD	Standard deviation
SDS-PAGE	Sodium dodecyl sulphate polyacrylamide gel electrophoresis
SOCS	Suppressor of cytokine signalling
SPOP	Speckle-type POZ
SR	Substrate receptor
SUMO	Small Ubiquitin-like Modifier
SUPREX	Stability of unpurified proteins from rates of H/D exchange
SVM	Support vector machine
TCEP	Tris(2-carboxyethyl)phosphine
TLR	Toll-like receptor
TM	Trajectory method
TOF	Time-of-Flight
TRAF	Tumour necrosis factor receptor-associated factor
TTR	Transthyretin
TWIMS	Travelling wave ion mobility spectrometry
UB	Ubiquitin
UBL	Ubiquitin-like protein
UI	User interface

UPLC	Ultra-performance liquid chromatography
UPS	Ubiquitin-proteasome system
UV	Ultraviolet
VBC	von Hippel Lindau-Elongin B-Elongin C
VHL	von Hippel Lindau-Elongin B-Elongin C
VMD	Visual molecular dynamics
WHA	Winged-Helix A
WHB	Winged-Helix b
WT	Wild-type
XL	Cross-linking

Chapter 1: Structural Mass Spectrometry for Modelling of Protein Complexes

1.1 Introduction to structural mass spectrometry

Proteins are diverse cellular machines that are capable of performing complex functions by virtue of their intricate structures. The intimate relationship between protein structure and function, can be studied through various biophysical and biochemical techniques that provide a method of interrogating their mechanical features, and inform on how it is that they fulfil their biological roles. While structural information can be accessible through techniques such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, these experiments are typically time consuming to perform, may have limitations in protein size and complexity, and produce models which are ambiguous in interpretation. X-ray crystal structures are such an example, whereby protein structures are stabilized in a solid lattice and often in a highly artificial and non-biologically relevant media. Although crystallographic structures offer unprecedented resolution into the intrinsic framework of proteins, further abnormalities in crystal contacts or crystal twinning, may complicate interpretations. Notwithstanding, the progressive shift in the focus of structural biology in the direction of protein conformational dynamics and interaction networks, means that new tools must be developed to fulfil new criteria in structural research.

Mass spectrometry is one of the most powerful analytical techniques. Its accuracy and sensitivity to detect atomic composition, from single atoms to intact viruses, megadaltons in mass, is unrivalled by other techniques. Hyphenation of MS with

other analytical methods such as ion mobility (IM), chemical cross-linking (XL) and hydrogen deuterium exchange (HDX) has evolved the MS technique into an attractive and practical toolkit for the study of biological molecules. While the resolutions of each of these techniques are lesser to that of crystallography, they each offer an alternative perspective on the molecular structure, composition and conformational behaviour of the system of study. Uniting data from these techniques along with existing crystallographic structures or other models, into streamlined computational workflows, has the potential to yield a more complete representation of a protein's dynamic capabilities. The following sections of this chapter will outline the principles of various concepts and MS techniques used throughout the work presented in Chapters 2, 3 and 4.

1.2 Overview of electrospray ionisation mass spectrometry

The analysis of biomolecules such as proteins and their complexes can be performed using mass spectrometry. At its core, the mass spectrometer consists of several components including an inlet, ion source, a mass analyser and a detector. At the inlet, a sample of protein must be subjected to ionisation in order to generate protein ions that can be detected via MS. One such method is via electrospray ionisation (ESI). ESI is a soft ionisation method which involves the gradual transfer of protein ions from an aqueous to gas environment (Figure 1.1).

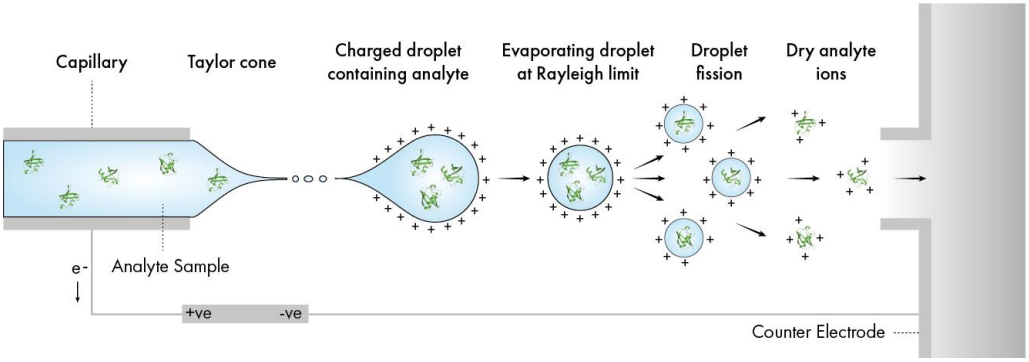


Figure 1.1. Analyte ionisation through ESI. Electrospray of an analyte sample from a capillary needle results in a spray of charged droplets containing analyte molecules. Desolvation of charged droplets results in charge accumulation, Coulombic repulsion and fission into smaller progeny droplets. Desolvation continues until only dry analyte ions remain and enter the mass spectrometer.

To achieve ESI, a potential difference is set up between a capillary holding the protein sample of interest (set to low microlitre min^{-1} flow rate) and an opposing counter electrode, establishing an electric field that draws the charged solvent from the capillary and towards the counter electrode. Charge accumulation at the capillary tip eventually causes instability of the solvent surface, leading to the formation of a Taylor cone which disperses the solvent into a fine plume of droplets roughly several microns in diameter[1]. Each droplet inherits a number of charges from the initial dispersion and are held together by the surface tension of the solvent.

The charge q that a droplet can hold in relation to its diameter D depends on the electrical permittivity of the surrounding environment ε and surface tension γ, given by the Rayleigh limit[2] **(1.1)**.

$$q = \sqrt{8\pi^2\varepsilon\gamma D^3} \qquad (1.1)$$

As droplets evaporate and decrease in size, the charge per unit volume increases until the Rayleigh limit is reached and Coulombic repulsion exceeds the surface tension of the droplet. This leads to droplet fission or ejection of charge from the surface of the droplet[3] (**Figure 1.1**). Desolvation continues until at some point, protein ions are generated (**Figure 1.1**). How ESI produces protein ions from desolvation is less certain, however for folded proteins, it is widely thought to occur under the framework of the charge residue model (CRM)[4]. The CRM envisions that protein ions are produced via gradual and complete desolvation of droplets. Under this mechanism, droplet desolvation continues until only single proteins are housed within each droplet. Evaporation of the final solvation shell is accompanied by the transfer of droplet charges to charge carriers on the surface of the protein[2,4,5]. Whether charge arises via protonation of basic (His, Lys, Arg) or neutralisation of acidic (Asp, Glu) residues is also a subject of discussion[6]. Several studies have further supported the notion that desolvation of proteins occurs under the CRM[3,7].

Following desolvation, dry protein ions enter the mass spectrometer and can be subject to different separation and fragmentation techniques to derive quantitative measurements on their molecular features. For example, the mass of an ion can be determined via *m/z* measurements using mass analysers such as quadrupole and time-of-flight (TOF) analysers. Features such as molecular shape can be quantified through ion mobility MS via measurements of collision cross-section using a collision cell.

1.3 Native mass spectrometry

Coupling ESI together with MS provides a method of studying macromolecules in a manner that presumably does not disrupt their three-dimensional structures. ESI of biological macromolecules under non-denaturing conditions is commonly referred to as "native MS". The exact suitability of the word "native" however, is controversial since MS techniques are undeniably performed in the gas phase[8]. It is widely accepted that the conditions of native MS are at least non-denaturing and ample examples of structural retention after transfer into the gas phase can be seen in native MS studies of proteins and their interactions with other proteins, small molecules, nucleic acids and lipids[9-12].

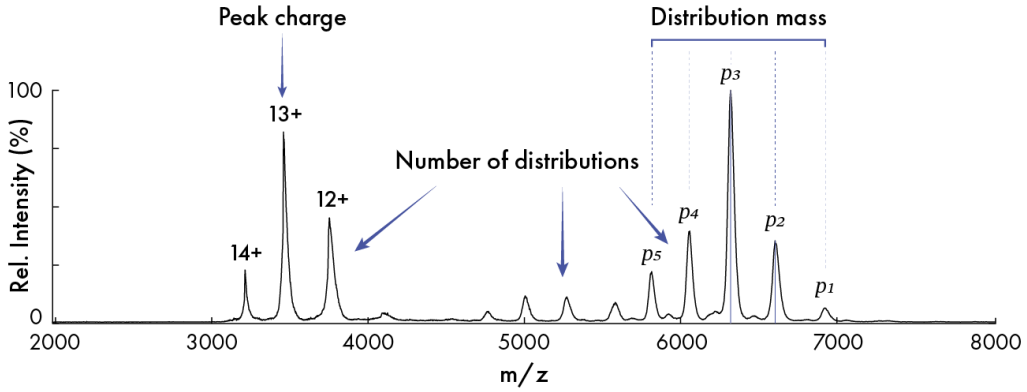


Figure 1.2. Example of a native MS spectra for a multi-subunit complex. Data shown is for the CRL2 complex (unpublished data).

MS is a separation technique - a protein sample subjected to native MS generates a mass spectrum that provides a readout of structural and compositional features. Since proteins and other macromolecules possess multiple charge carrying sites, ionisation via ESI produces a statistical distribution of multiply charged ions (Figure 1.2). Samples containing multiple molecules will produce multiple distributions of peaks (Figure 1.2). Each distribution corresponds to a single molecular population with mass m, while each constituent peak p corresponds to one of its observed charge states z, and thus manifests as a distribution of *m/z* values. The highest *m/z*

peak of a distribution corresponds to the lowest charge state z_1 and the charge of each subsequent adjacent peak is given by $z_n = z_{n-1} + 1$. For two adjacent peaks p_1 and p_2, their *m/z* values can be written as (1.2):

$$p_1 = \frac{m + zH^+}{z} \qquad p_2 = \frac{m + (z+1)H^+}{z} \tag{1.2}$$

where H^+ is the mass of a proton at 1.008 Da. Since both m and z are unknown, calculating these values from p_1 and p_2 involves solving a pair of simultaneous equations. p_1 and p_2 can be made equal by rearranging for m (1.3):

$$zp_1 - zH^+ = \ z+1\ p_2 - \ z+1\ H^+ \tag{1.3}$$

Rearranging for z gives (1.4):

$$z = \frac{p_2 - H^+}{p_1 - p_2} \tag{1.4}$$

Substituting the spectral *m/z* values for p_1 and p_2 calculates z of p_2 which can be re-substituted into (1.2) to derive the m of p_2. In practice, mass calculations via the above methods can be performed in an automated manner using MS software such as MassLynx (Waters Corp.). While mass is no doubt the most important parameter assessed by native MS, Hall & Robinson have demonstrated that the average observed charge state of a distribution can additionally be used to predict the corresponding solvent accessible surface area (SASA) of a molecule in good agreement with calculated SASA from crystal structures (Figure 1.3)[13].

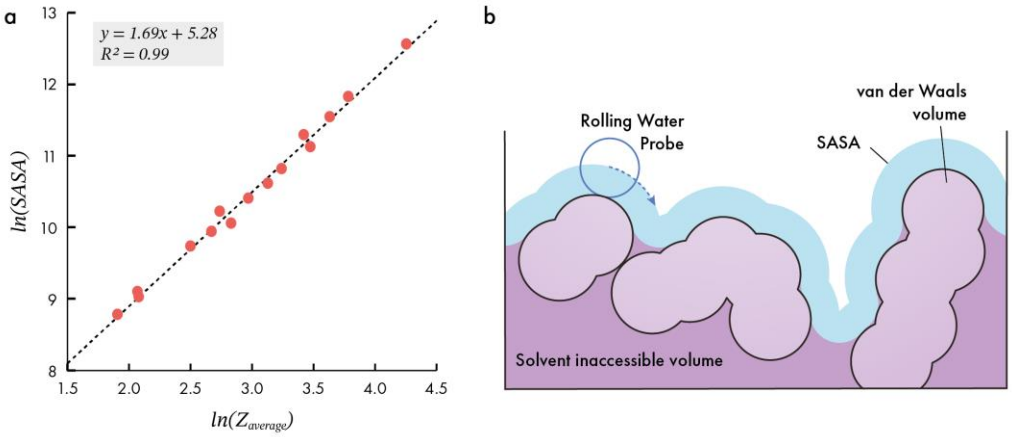


Figure 1.3. Relationship between charge from native MS and SASA. (a) Empirical relationship between the average observed charge state $Z_{average}$ and SASA. Adapted from Hall & Robinson, 2012 [13]. (b) SASA is defined as the surface area accessible by a water molecule (modelled by a spherical probe of radius 1.4 Å).

1.4 Ion mobility mass spectrometry

Using native MS, populations of ions are separated via differences in their *m/z*. Hyphenation of native MS with ion mobility (IM-MS) provides a method of further separating ions of the same *m/z*, based on their shape[14,15]. The quantitative measurement derived from IM-MS is the collision cross-section (*CCS*) - a physicochemical feature of the ion's size, shape and charge[16]. Differences between the conformations of ions of the same *m/z* manifest as differences in their *CCS*, allowing subpopulations of molecular shapes to be distinguished between. The earliest methods of IM-MS utilised linear drift tubes (DT) in which analyte ions are propagated through an inert gas (typically He or N_2) filled chamber of pressure p, by a weak electric field E (Figure 1.4).

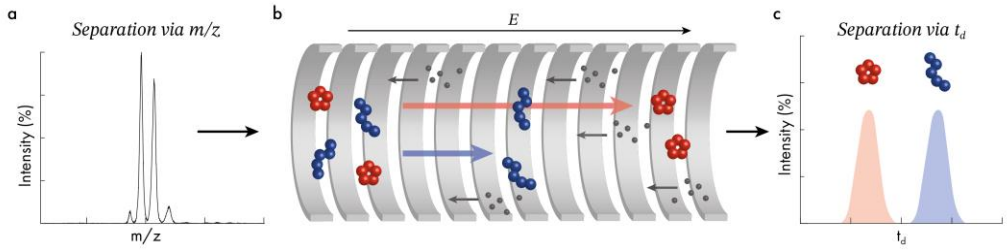


Figure 1.4. Ion mobility allows further separation of ions of same *m/z* according to their molecular shape. (a) Example of native spectra indicating presence of a single population of ions. (b) Inert gas molecules (grey spheres) continuously collide with analyte ions (red and blue spheres) which are propelled through the collision cell via an electric field ($\boldsymbol{E}$). Any differences between the mobility of the ions, such as their *CCS*, lead to differences between the time taken for the ions to cross the collision cell. (c) The corresponding mobility spectra from (b) distinguishes between two separate distributions of molecular shapes.

Sub-populations of ions with a more compact profile between two populations of otherwise identical ions, will have greater mobility K and thus traverse through the chamber quicker due to fewer collisions with the buffer gas compared to a more elongated conformation. K is defined as the ratio between the steady state velocity of the ion v_d, and the applied electric field (1.5). In turn, the velocity is the ratio between the length of the drift tube l and the time taken to traverse it t_d (1.5).

$$v_d = \frac{l}{t_d} \qquad K = \frac{v_d}{E} = \frac{l}{t_d E} \tag{1.5}$$

As K is dependent on the number of collisions experienced by the ion and is propagated by the electric field, it is proportional to E and inversely proportional to both p and the number of gas particles in the chamber N and is affected by the temperature T. The *CCS* of an ion is calculated from K using the Mason-Schamp equation (1.6):

$$CCS = \frac{3}{16}\sqrt{\frac{2\pi}{\mu k_B T}}\frac{ze}{N}\frac{1}{K} \tag{1.6}$$

where μ is the reduced mass of the ion and gas pair, k_B is the Boltzmann constant, z is the absolute charge of the ion and e is the elementary charge. To ensure the consistency of *CCS* measurements across laboratories, K is corrected to standard

pressure p_0 (100,000 Pa) and temperature T_0 (275.15 K) and reported as reduced mobility K_0[17]. N is also corrected to the Loschmidt's constant N_0 (1.8).

$$K_0 = K \cdot \frac{N}{N_0} = K \cdot \frac{p_0}{p} \cdot \frac{T}{T_0} \tag{1.7}$$

$$CCS = \frac{3}{16}\sqrt{\frac{2\pi}{\mu k_B T}}\frac{ze}{N_0}\frac{1}{K_0} \tag{1.8}$$

In summary, for a simple linear drift tube system, the time taken for an ion to traverse the collision chamber t_d is used in (1.5) to calculate the mobility of the corresponding ion. K is corrected for measurement-specific parameters such as N, p and T in (1.7) and then used to calculate *CCS* in (1.8).

In commercial instruments such as the Synapt G2-Si (Waters Corporation) with which some data shown in this thesis were acquired using, IM-MS measurements are performed using travelling wave or "T-wave" IM-MS (TWIMS). Unlike the drift tube technique which measures t_d of an ion in a linear trajectory, TWIMS employs a periodic travelling wave of specific wave height and velocity, which increases the separation of ions but does not allow *CCS* to be determined directly from t_d (Figure 1.5). Instead, the t_d of calibrant proteins which have had their *CCS* measured via drift tube, is used to calibrate the t_d of the analyte to derive its equivalent *CCS* (Figure 1.5).

To derive *CCS* using TWIMS, the time of arrival at the end of the collision chamber t'_d is measured for each charge state of multiple calibrants. Which calibrants are used, depends on the analyte of interest and should be selected such that analyte mass is between that of the calibrant masses[18]. Automated calibration can be performed using software such as PULSAR[18]. The t'_d of each calibrant is plot against its expected drift tube measured *CCS*, CCS' allowing a calibration curve to be generated and applied to the analyte data.

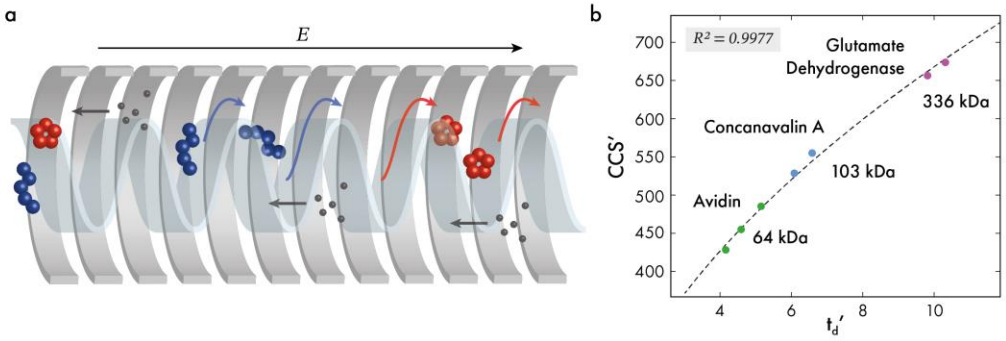


Figure 1.5. Separation of ions using TWIMS. (a) The separation of ions can be improved in IM-MS by applying a travelling wave of specific wave height and velocity. (b) Example calibration curve generated from TWIMS measured t'_d of avidin, concanavalin A and glutamate dehydrogenase and plotting against literature *CCS* values determined via drift tube measurements (CCS'; *unpublished data courtesy of Z. Ahdash*). The approximate molecular weight of each calibrant is shown as well as the R^2 of the fitting. Individual datapoints for each calibrant represents an alternative charge state.

1.5 Methods of calculating theoretical *CCS*

Experimental measurements of *CCS* can be compared with theoretical values calculated from atomistic or coarse-grained models of proteins and constitutes the basis of molecular modelling using data from IM-MS. Many methods have been conceived in order to calculate theoretical *CCS* from models. These can be subdivided into two groups: methods that perform scattering calculations to derive *CCS*, including the exact hard sphere scattering (EHSS) and trajectory method (TM), and those that approximate *CCS* through non-scattering calculations such as projection approximation (PA) and projected superposition approximation (PSA). The features of each method will be briefly described below.

The earliest examples of comparisons between experimental and theoretical *CCS* can be found from almost a century ago by Mack, who in 1925, envisioned that model *CCS* could be approximated through averaging an object's projected area over all rotational angles[19]. The method of calculating model *CCS* based on its projected area were later encapsulated in the projection approximation (PA)

method. Measurements of PA along with EHSS and TM can be performed using software such as MOBCAL by Mesleh *et al*.[20]. Measurements of *CCS* via PA (CCS_{PA}) takes into account the size of the buffer gas used, but not any explicit interactions between protein and gas particles since scattering is not modelled using PA. The projected area is calculated for a number of different orientations of the molecule however non-globular topologies which present surfaces for multiple scattering events are typically under-represented, leading to underestimation of the true *CCS*[14,21]. However as demonstrated by Benesch and Ruotolo, an empirical scaling factor of 1.14 can be applied to yield *CCS* that is in good agreement with experimental equivalents[22]. Calculations of *CCS* via PA are among the fastest of all methods (Table 1.1). Recent upgrades to the PA calculation method by Marklund *et al*. in the software IMPACT, has made it possible to derive CCS_{PA} for all biological assemblies in the protein data bank (~300,000 models) in approximately 5 hours[23], making PA a highly viable *CCS* determination method when a large number of models must be accounted for.

Limitations in the PA to account for some macromolecular shapes such as concave surfaces, are further remedied by the PSA method which includes size and shape effects, allowing more representative *CCS* values to be calculated for non-globular shapes such as channels, cavities and pores[24]. PSA approximates long range interactions via application of statistical probabilities to take into account temperature and distance-dependence of scattering events. As a result, calculations with PSA are much longer than those of PA[14].

Table 1.1. Methods of calculating *CCS* from models.

Method	Typical computing timescale
PA	< seconds
PSA	minutes
EHSS	minutes
TM	hours

More rigorous calculations of *CCS* can be made through EHSS and TM methods, both of which calculate *CCS* as a momentum transfer integral, taking into account the sum of individual collision events. The most accurate calculations of *CCS* are derived from the TM which performs repeated scattering simulations of the model of interest colliding with gas particles while taking into account both short- and long-range interactions[20]. TM integrates the momentum-transferring collision events over all scattering orientations simulated and uses molecular mechanics to simulate a realistic trajectory of the gas molecule[20,25]. The high number of calculations performed for both short- and long-range interactions means that the TM is very computationally expensive. TM for protein molecules can take several hours to compute, indicating that the accessibility of CCS_{TM} is severely limited by the number of models or system size. A more simplistic method of modelling the scattering trajectories is offered by the EHSS method, which speeds up calculation times but at the cost of ignoring long range interactions between gas and model particles[26].

A new and unique approach to calculating *CCS* of models emerged in 2016 from Zhou *et al*. who used a support vector machine (SVM) to predict the model *CCS* of small molecules based on 14 molecular features, including mass, formal charge, physiological charge and solubility[16]. SVMs are a subset of artificial neutral networks (a branch of machine learning) which can be trained to partition data based on a set of descriptive parameters (in this case the molecular features) and an expected result (the *CCS*). A hyperplane is constructed through the datapoints and used in regression analysis to derive a relationship between the input descriptive parameters and output expected *CCS*. Zhou *et al*. applies their SVM to the human metabolome database of 35,000 compounds and show accurate *CCS* derivation for 90% of the compounds[16]. *CCS* via machine learning has so far only been applied to small molecules and no algorithms have been trained for protein complexes. However, the application of machine learning in the field of structural mass spectrometry is an interesting method of empirically modelling relationships between different structural and chemical features of molecules.

1.6 Collapse of protein structures in the gas phase

As seen previously, a combination of native and IM-MS can distinguish between sub-populations and sub-conformations of protein structures through differences in their *m/z* and *CCS*. Whether these structures resemble their solution counterparts is a vigorously discussed topic and highly relevant when considering the possible implications of misinterpreted structural information. The consensus that ESI produces proteins that remain folded upon transfer into the gas phase, is supported by a number of observations including: (1) protein spectra following native MS do not resemble those of unfolded proteins which are able to carry much more charge, and thus suggest compact gas phase topologies; (2) native MS charge states correlate with calculated SASA from protein models, further supporting (1); (3) non-covalent interactions between protein-protein or protein-ligand are preserved in the gas phase, suggesting that the gas phase topology resembles that of the solution phase; (4) experimental *CCS* of most proteins can be well correlated with theoretical values calculated from crystal structures. Point (4) is especially interesting and worthy of further discussion since a number of proteins and nucleic acids have been observed, whereby their experimental *CCS* can be up to 60% lower than that of the theoretically calculated value (Table 1.2, Figure 1.6).

Jurneczko and Barran review the *CCS* of number of proteins shown in Figure 1.7, and observed that for all examples, experimental *CCS* is consistently lower than that of the theoretical value[21]. Ewing *et al*. further report that theoretical *CCS* calculated via TM, although being one of the more rigorous models of collision events, typically results in a +5% *CCS* difference compared to experimental values[27]. To explain these differences, Jurneczko and Barran rationalise that protein-solvent interactions are important for structural stability in the solution environment, and so protein structures "collapse" during transfer into the gas phase as evaporate to dryness[21].

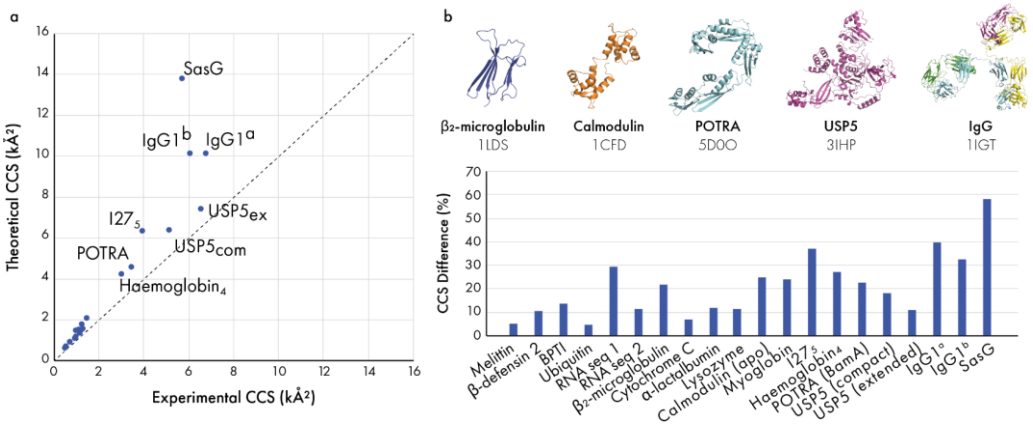


Figure 1.6. Differences between theoretical and experimental *CCS* found in the literature. (a) Plot of theoretical and experimental *CCS* for 20 molecules. Points closest to the origin have not been labelled for clarity. Dashed line represents exact match between experimental and theoretical values. (b) Percentage *CCS* difference between datapoints in (a) are shown, along with five example protein structures. Datapoints correspond to those displayed in **Table 1.2**. Abbreviations: $Haemoglobin_4$: haemoglobin tetramer; $USP5_{com}$ and $USP5_{ex}$: compact and extended USP5; BPTI: bovine pancreatic trypsin inhibitor. $IgG1^a$ values from Devine *et al*. 2017 [28]. $IgG1^b$ values from Pacholarz *et al*. 2014 [29].

Table 1.2. Reported experimental and theoretical *CCS* values

Protein	MW (Da)	PDB	Method	CCS_{exp}	CCS_{Thr}	Method	%
Melittin[a]	2846	1MLT	DT	544	574	TM	5.2
β-defensin 2[a]	4328	1LFD3	DT	598	669	TM	10.6
BPTI[a]	6512	6PTI	DT	770	891	TM	13.6
Ubiquitin[a]	8565	1UBQ	DT	1004	1055	TM	4.8
β2-microglobulin[a]	11860	1LDS	TWIMS	1142	1459	TM	21.7
Cytochrome C[a]	12355	1HRC	DT	1217	1310	TM	7.1
α-lactalbumin[a]	14178	1HFX	TWIMS	1342	1523	TM	11.9
Lysozyme[a]	14305	1DPX	DT	1300	1468	TM	11.4
Calmodulin (apo)[a]	16700	1CFD	DT	1526	2029	TM	24.8
Myoglobin[a]	17566	1VXG	TWIMS	1314	1725	TM	23.8
Haemoglobin[a]	64447	1GZX	DT	3051	4181	TM	27.0
USP5[b]	93792	3IHP	TWIMS	5200[c]	6340	TM[d]	18.0
			TWIMS	6600[c]	7400	TM[d]	10.8
IgG1[e]	150000[f]	1IGY	TWIMS	6820	10100	PSA[g]	32.5
$I27_5$[e]	52235[h]	N/A[i]	TWIMS	3980	6310	PSA[g]	36.9
POTRA (BamA)[e]	90552[j]	5D0O	TWIMS	3510	4530	PSA[g]	22.5
SasG[e]	178527	N/A[i]	TWIMS	5770	13780	PSA[g]	58.1
RNA (seq 1)[e]	11217	2PCV	TWIMS	1021	1445	PSA[g]	29.3
RNA (seq 2)[e]	11219	2DRB	TWIMS	1016	1146	PSA[g]	11.3
IgG1[k]	150000[f]	1IGY	DT	6100	10084	TM	39.5

a Values from Jurneczko & Barran, 2011 ref 21

b Values from Scott, Layfield and Oldham, 2015 ref 30

c compact and extended conformations of USP5

d Approximate TM-like value via $CCS_{TM} = PA \times 1.14$

e Values from Devine *et al*., 2017 ref 28

f Exact mass not given

g Approximate PSA-like value via $CCS_{PSA} = \ PA - 81\ \times 1.299$

h Mass from Brockwell *et al*. 2002 ref 31

i Modelled by authors, no model deposited

j Mass from UniProt (AN: P0A940)

k Values from Pacholarz *et al*., 2014 ref 29

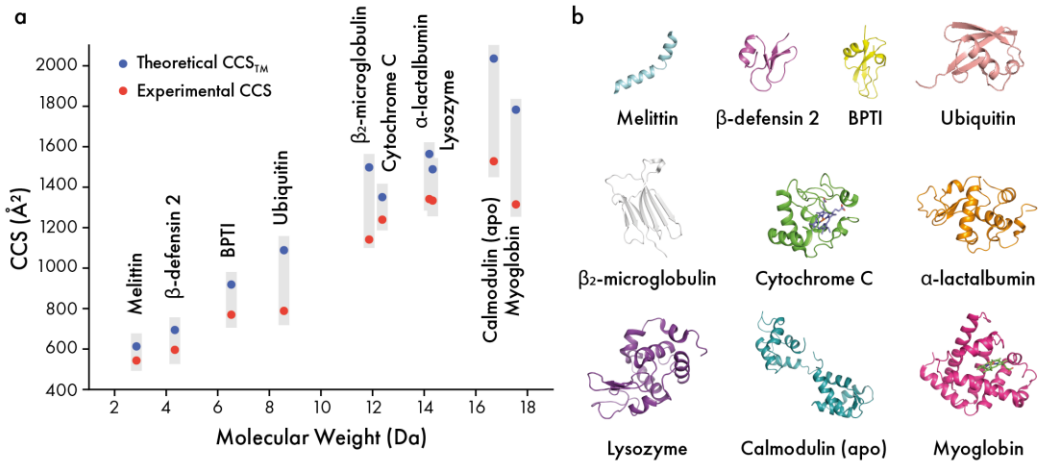


Figure 1.7. Comparison of theoretical and experimental *CCS* for proteins. (a) The experimental *CCS* of melittin, β-defensin 2, bovine pancreatic trypsin inhibitor (BPTI), ubiquitin, β2-microglobulin, cytochrome C, α-lactalbumin, lysozyme, apo conformation of calmodulin and myoglobin are shown along with theoretical *CCS* calculated via TM. (b) Crystal structures used to derive values in (a). Adapted from Jurneczko and Barran, 2011 [21].

A model of the timeline following desolvation was described by Breuker and McLafferty who used molecular dynamics (MD) simulations of cytochrome C by Steinberg *et al*. to propose that desolvation eventually leads to the loss of protein structure over millisecond timescales[32]. In this model, charged protein side chains collapse within picoseconds of complete desolvation, forming a stabilising salt-bridge network that encapsulates the surface of the protein[32]. Devoid of its solution environment, loss of the hydrophobic effect follows, eventually leading to a loss of salt-bridges and unfolding of cytochrome C over the next few milliseconds[32]. However, drift time measurements of ubiquitin by Koeniger, Merenbloom and Clemmer, demonstrated that the gas phase conformation of ubiquitin does not change or unfold over the 10-20 ms timescale of the experiment[33], indicating that stability should be discussed on a per-protein basis.

Further attempts to model gas phase conformations of proteins, as well as the desolvation process have provided valuable insights into the relationship between protein topologies and gas phase collapse. Methods of simulating these events through molecular dynamics will be discussed later. Gas phase collapse typically occurs in proteins with non-globular topologies, particularly those with domains

connected by flexible linkers[28]. Compact gas phase structures of extended and elongated proteins such as the IgG1 antibody, and two rope-like proteins known as $I(27)_5$ and SasG, have been simulated. However, the post-MD theoretical *CCS* values of these simulations typically still exceed that of the experimental *CCS*[28,29]. On the other hand, gas phase simulations of proteins with fewer flexible linkers such as the USP5 deubiquitinase and the POTRA domains of the BamA complex, have yielded structures highly matching experimental *CCS*[28,30]. The observation that not all simulations may be able to produce agreeable models, suggests that either the simulation method or the initial starting conformations of proteins, typically found from crystal structures, may be limited in their ability to represent the solution ensemble and may lead to structure being trapped in non-representative conformations. Modelling successes of proteins with fewer degrees of freedom such as USP5 and POTRA, support this notion further and indicate a correlation between the in-solution conformational space of a protein, and its degree of gas phase collapse. The methods of performing gas phase simulations on protein structures will be explored in the following section.

1.7 Principles of molecular dynamics simulations

The motions of proteins in the form of their dynamic interactions with itself, or atoms in the surrounding environment are an essential feature of its conformational and biochemical space. MD simulations provides a method of simulating the motions of proteins and their associated mechanical events as a function of time[34]. The availability of high-performance computing power, graphical processing units and developments into accurate parameterisation of MD forcefields have led to the ability to simulate increasingly complex and diverse biological systems to even millisecond timescales[35]. Several variants of MD simulations exist, including classical and quantum mechanical methods. MD performed for macromolecular motions are typically referred to as "classical MD" - differentiating them from simulations such as enzymatic reactions which require quantum mechanical methods that are able to model the breaking and forming of covalent bonds to high accuracy.

In classical MD simulations, the trajectory of each atom is calculated through integrating Newton's second law over small timeframes, typically femtoseconds. Each calculation step of the trajectory involves calculating the force acting upon each atom according to its environment. The atom's position is then updated, and the process repeated. The physical characteristics of atoms during MD are represented in a forcefield that is applied globally to all atoms. These include the atomic mass, partial charge, van der Waals radii, and inter-atomic bond lengths and angles. One of the most widely used MD forcefields for the simulation of biological molecules is the CHARMM[l] forcefield developed by Martin Karplus and colleagues[36]. Many flavours of the CHARMM forcefield have been developed, each specialising in simulations of a particular type such as the CHARMM27 for protein-nucleic acid

l CHARMM: Chemistry at Harvard Macromolecular Mechanics

interactions[37,38]. In the CHARMM forcefield, the energy E for a set of protein atoms $\mathbf{R}$ is calculated using an additive six-term function (1.9):

$$E(\boldsymbol{R}) = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{dihedral} k_\phi[1 + cos(n\phi - \delta)] + \sum_{improper} k_\omega(\omega - \omega_0)^2 + \sum_{UB} k_u(u - u_0)^2 + \sum_{nonbonded} \varepsilon\left[\left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{12} - \left(\frac{R_{min_{ij}}}{r_{ij}}\right)^{6}\right] + \frac{q_i q_j}{\varepsilon r_{ij}} \quad (1.9)$$

Each of the six terms describe different bonded (terms 1-5) and non-bonded (term 6) interactions (Figure 1.8). The first term in (1.9) calculates the potential energy contribution of bond displacements b from an ideal bond length b_0, where k_b is the bond force constant. Both b_0 and k_b are parameterised for each bond type (e.g. C-C, C-H, N-H). In the second term, oscillations around bond angles are also modelled via a harmonic potential, and k_θ is the corresponding angle force constant for the particular set of atoms.

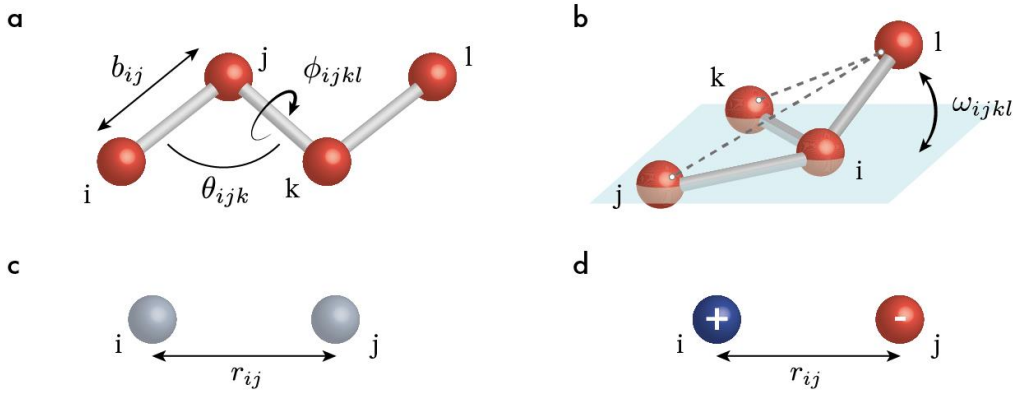


Figure 1.8. Bond characteristics. (a) Definitions of bond lengths $\boldsymbol{b}$, angles $\boldsymbol{\theta}$ and dihedrals $\boldsymbol{\phi}$ for a set of four bonded atoms. (b) Improper dihedrals (out-of-plane) are measured as the angle between planes $\boldsymbol{ijk}$ and $\boldsymbol{jkl}$ and given as $\boldsymbol{\omega_{ijkl}}$. (c-d) van der Waals and electrostatic interactions between atoms $\boldsymbol{r_{ij}}$ distance apart.

Dihedrals of a set of four atoms are assumed to be periodic and taken as the sum of cosine functions where n is the multiplicity, ϕ is the dihedral angle and δ is the phase shift. k_ϕ is the corresponding atom-specific dihedral force constant for the set of four atoms that comprise the dihedral angle. Out-of-plane or "improper" dihedrals are modelled similar to terms 1 and 2 as harmonic potentials of the angle between two atom planes shown in Figure 1.8. The 5th term known as the Urey-Bradley potential, is specific to the CHARMM forcefields and introduces a virtual bond between two atoms involved in an angle and restrains the distance between them u, based on idealised values u_0 and an atom-specific force constant k_u. Finally, contributions from non-bonded interactions are taken as the sum of van der Waals and electrostatic bonds and calculated for atom pairs separated by no more than three bonds (1.10):

$$E_{nonbonded} = E_{vdw} + E_{electrostatic} \tag{1.10}$$

$$E_{vdw} = \varepsilon \left[\left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{12} - \left(\frac{R_{min_{ij}}}{r_{ij}} \right)^{6} \right] \qquad E_{electrostatic} = \frac{q_i q_j}{\varepsilon r_{ij}} \tag{1.11}$$

van der Waals interactions are modelled using a Lennard-Jones potential where $R_{min_{ij}}$ is the atom type-dependent distance at which the potential energy is zero, r_{ij} is the distance between atoms i and j and ε is the dielectric constant for the surrounding media (1.11). The electrostatic energy term is represented by a Coulomb potential taking into account the product of atom charges $q_i q_j$, ε and distance between them r_{ij} (1.11).

1.8 Molecular dynamics of proteins in the gas phase

MD simulations that aim to elucidate functional aspects of proteins are typically performed in a solution environment that mimics its native aqueous environment. The behaviour of water molecules and protein solvation effects are handled by a number of different water models[39]. MD simulations are also commonly combined

with structural MS studies in order to leverage MS observations with atomistic representations of the system of study. This has led to fruitful discoveries including contributions into the understanding of how DNA is repaired by the HerA-NurA complex which was performed during this PhD but is not shown in this thesis[9]. Marklund and Benesch provide an excellent review of synergy between MD and MS[40].

MD simulations of proteins in a vacuum environment are commonly performed and used in conjunction with IM-MS. However, due to the lack of forcefields developed for proteins in the gas phase, those for solution environments are typically used in their stead, leading to controversy regarding the validity of the resulting structures. The effectiveness of performing gas phase simulations with solution forcefields has been described by Konermann *et al*.[41]. Konermann and colleagues rationalise that there are multiple criteria that a forcefield developed for the gas phase must take into account, primarily a reasonable representation of the atoms at the protein-vacuum interface and also at the protein interior which would likely be better represented by condensed-phase forcefields[41]. Notwithstanding the lack an ideal gas phase forcefield, Konermann *et al*. point out that so far, gas phase simulations have yielded "surprisingly good" descriptions of proteins in the gas phase but conclude that all simulations should be judged on their ability to provide new insights into the system of study.

Current methods of performing gas phase MD of proteins can be divided into those that include the timeline of desolvation and those that do not (Figure 1.9). To model the ESI process, the group of Lars Konermann developed a trajectory stitching method and demonstrated its utility in simulating the desolvation of ubiquitin, cytochrome C and holo-myoglobin using the CHARMM36 forcefield developed for proteins in solution[42]. In these simulations, water droplets each containing a single molecule of protein is simulated over a period of >100 ns and droplet charges are modelled through an excess number of Na^+ (Figure 1.9a). Water molecules evaporate from the droplet, accompanied by ejection of charge, as envisioned under the framework of ESI. Inspecting the time evolution of the droplet charge also indicated

that Na^+ ejection occurred as the droplet charge approached the Rayleigh limit[41]. To lessen the computational cost, the simulation is periodically stopped, water molecules distanced greater than 150 Å from the protein centre-of-mass are removed, and the simulation continued. Eventually, each simulation results in a dry $[M + zNa]^{z+}$ protein ion with a theoretical *CCS* in good agreement with experimental measurements[42]. The agreeability of McAllister's simulations to capture key events of the ESI process using the CHARMM36 forcefield, supports Konnerman's rationale that solution forcefields although not ideal, may be equally valid for gas phase simulations.

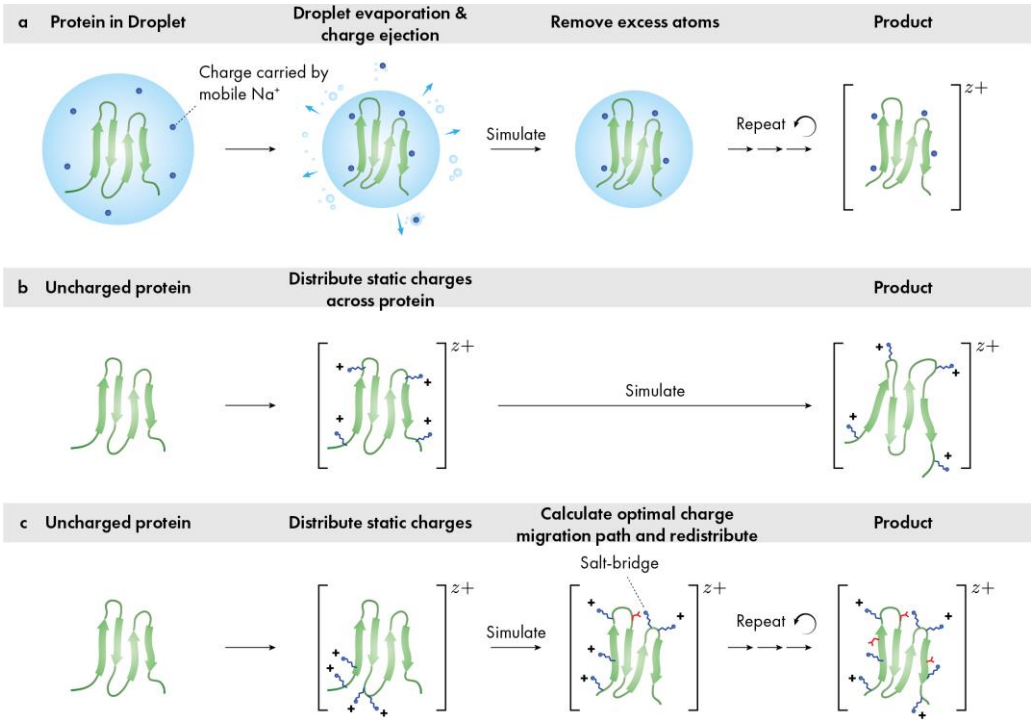


Figure 1.9. Methods of protein gas phase MD. (a) Simulating protein desolvation involves placing a protein of interest in a water droplet. Charges are contributed by excess ions such as Na^+. After a short period of simulation, all atoms exceeding 150 Å from the protein centre-of-mass are removed and repeated until the protein is dry. (b) Proteins can be directly subjected to gas phase MD. This involves pre-charging the protein through distributing charges across the surface of the protein, typically to basic residues. All other residues are assumed to be neutral. The protein is simulated until convergence. (c) Similar to (b), an initial distribution of static charges are distributed across the protein surface. After a short period of simulation, an optimal migration path for each charge is calculated. Charges are moved to the new residues if more energetically favoured. Another short simulation takes place and the process is repeated until convergence. Charges are further allowed to migrate from acidic residues, allowing salt-bridges to be formed.

There are equally examples where proteins have been subjected directly to gas phase MD and these constitute the majority of reports found in literature[28-30]. Unlike desolvation simulations which model droplet charge and protein ionisation using mobile Na^+, direct gas phase protein simulations first perform pre-charging of the protein prior to simulation (Figure 1.9b)[41]. Typically, all residue side chains are neutralised such that the net protein charge is zero. A set number of charges (those seen from ESI) are distributed via protonation of either a randomised or calculated distribution of basic residues such as His, Arg and Lys. Another limitation manifests in this step as not all forcefields contain parameters for protonated residues. In these simulations, a limitation of MD engines such as GROMACS or NAMD to handle mobile charges, means that the initial charge setup is fixed over the course of the simulation. As a consequence, the initial charge set up may heavily bias the resulting trajectory. Meyer *et al*. rationalise that charging a small protein such as lysozyme to 5+, with 19 basic Arg, Lys, His and N-termini sites, results in more than 11,000 possible combinations to consider[43]. A mobile proton algorithm was developed by Konermann and colleagues to circumvent these major simplifications in the pre-charging method[44]. The mobile proton method uses the OPLS forcefield designed for solution simulations. In essence, the algorithm first performs a pre-simulation sampling of thousands of charge distributions and determines the most energetically stable selection of chargeable sites including this time Asp, Glu, N and C-terminal sites[41]. In a manner similar to that of the trajectory stitching method, the algorithm performs short stretches of simulations followed by re-calculation and re-distribution of charges (Figure 1.9c). As a result of mobile charges and inclusion of both acidic and terminal residues, salt bridges are abundantly formed during mobile proton simulations[44]. Both simulations using the trajectory stitching desolvation and mobile proton methods are detailed in a protocol paper by Konermann *et al*.[41]. These new methodologies will no doubt lead to an increasingly accurate and informative number of gas phase simulations in the future.

1.9 Chemical cross-linking mass spectrometry

Synonymous to the utility of MD in providing access to the dynamics of protein structures, their conformational dynamics and interaction space can be captured experimentally through chemical cross-linking (XL)-MS. XL-MS provides proximity data, and thus when applied to proteins and their complexes, can be used to determine changes in protein conformations, the proximity of regions within multi-subunit complexes and also determine interaction networks in proteome-wide studies[45]. XL-MS combines together the structural specificity of small chemical modifiers with the analytical sensitivity of MS. These chemical modifiers, known as cross-linkers, are bifunctional molecules that feature two reactive groups separated by a linker or spacer of defined length. The reactive groups of a cross-linker are specially designed such that they react with particular functional groups of proteins or other proximal interacting molecules such as nucleic acids or lipids[46,47].

A typical XL-MS workflow involves first incubating the system of interest with cross-linker molecules (Figure 1.10a-b). There are several types of cross-linking experiments, each with a range of cross-linkers including amine-to-amine, zero-length and photo-activated cross-linkers (Figure 1.11). Cross-linkers can further be homo- or hetero-bifunctional, giving rise to a diverse palette of chemical specificities (Figure 1.11c)[45]. Amine-to-amine cross-linkers, such as the commonly used bis(sulfosuccinimidyl) suberate or BS3, is a homo-bifunctional molecule with two N-hydroxysuccinimide (NHS) esters separated by an 8-carbon long spacer for a total length of 11 Å. When exposed to proteins, the NHS groups of BS3 react with amine groups of Lys and the N-termini, although they have also been shown to have limited reactivity for hydroxy groups of Ser, Thr and Tyr[48].

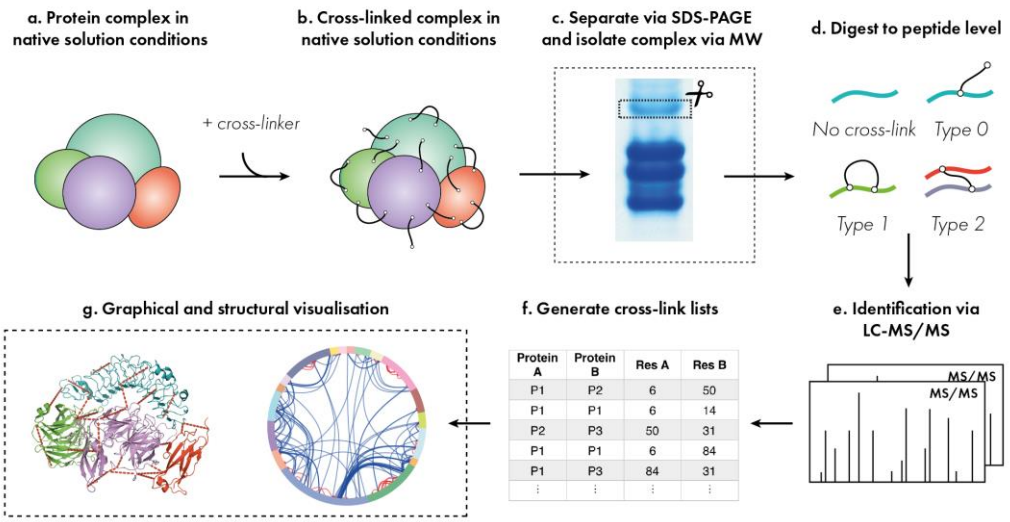


Figure 1.10. Example of the XL-MS workflow. (a-b) Protein complexes in an aqueous environment are incubated with cross-linker reagents which tether together pairs of residues within a defined distance. (c) An optional step involves the fractionation and isolation of specific MW species via SDS-PAGE. (d) Cross-linked proteins are digested to the peptide level using various enzymes (e.g. trypsin). Four types peptides can be generated: non-cross-linked peptides, type 0 mono-links in which one reactive group of the peptide has not reacted, type 1 cross-links where both groups are found tethered to the same peptide, and type 2 cross-links where both groups are found on different peptides. (e) The cross-linked peptide mixture is subjected to LC-MS/MS and database search for cross-link identification. (f-g) Tables of cross-linked residues are generated and can be visualised either graphically or via projection onto existing or modelled structures of the system. Steps (c) and (g) are optionally performed.

Carbodiimides, condense together carboxylate and amine groups of proteins, forming zero-length linkages. Other cross-linkers such as diazirines are photo-activatable[49]. Under long wave UV irradiation (330-370 nm), diazirines form reactive carbene intermediates which have non-selective reactivity for solvent accessible surfaces of the protein[45]. Incorporation of diazirine-modified photo-activatable amino acids such as photo-lysine have even enabled the capturing of protein-protein interactions directly from within cells[50]. In the case of a bifunctional cross-linker such as BS3, the reactive groups engage with amine groups of the protein, stabilising its structure via a network of covalent linkages (Figure 1.11a).

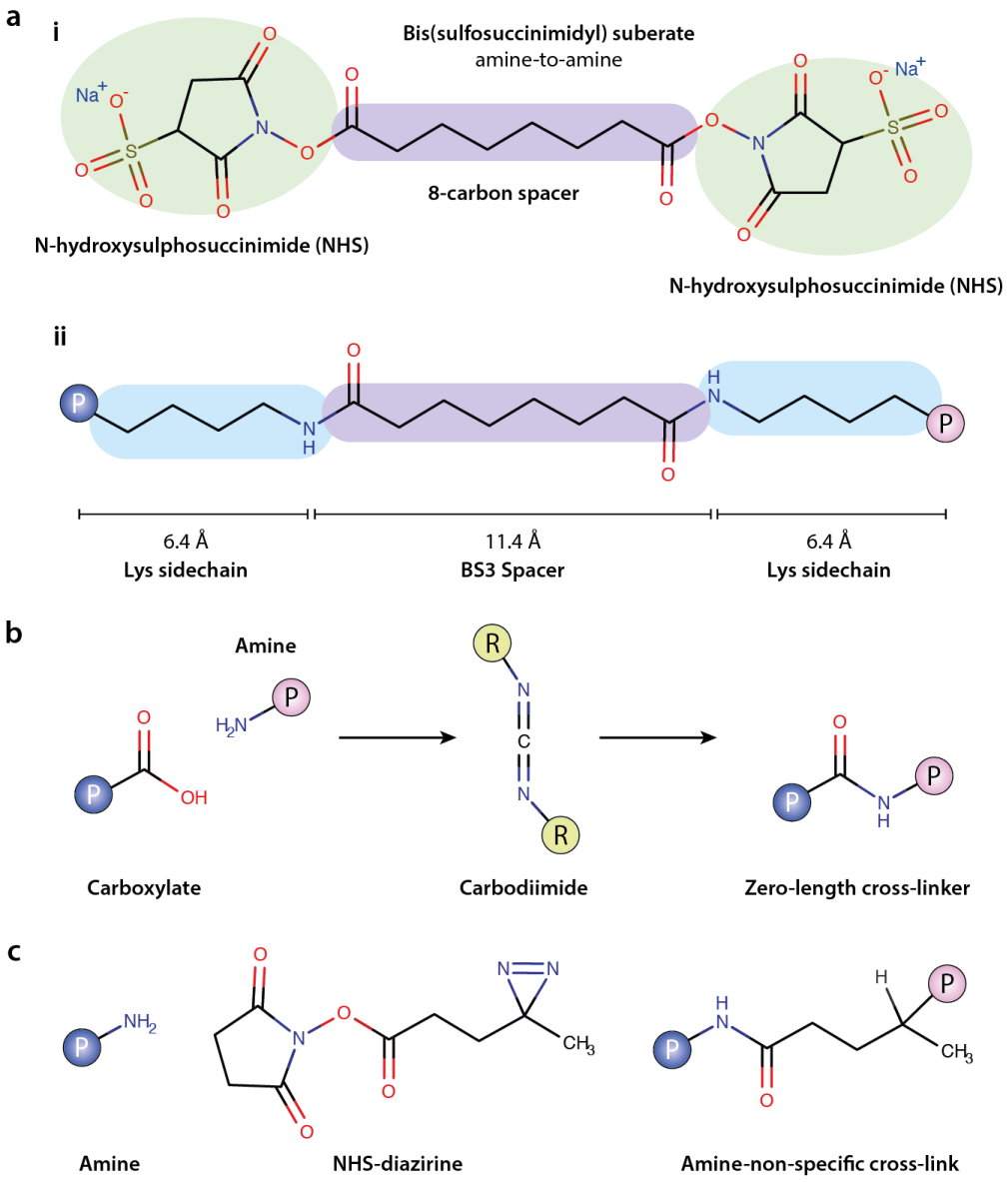


Figure 1.11. Cross-linker chemistries. (a) Amine-to-amine cross-linking using BS3. (i) BS3 consists of two amine-reactive NHS groups separated by an 8-carbon spacer. (ii) Cross-linked lysine side chains using BS3. (b) Carboxylate and amine groups of proteins can be zero-length cross-linked using carbodiimides. (c) Hetero-bifunctional cross-linkers such as NHS-diazirine are bi-specific. NHS reacts with amine groups while the diazirine upon UV irradiation, generates a reactive carbene species that non-specifically reacts with proximal chemical groups.

A major benefit of XL-MS is that cross-linking events take place in solution, allowing structural and conformational information to be captured from proteins while in an

environment that is most representative of its conformational dynamics. Since cross-linking takes place in solution, the data acquired from XL-MS are ensemble measurements that represents all sub-populations and sub-conformations of protein states present within the sample. This presents multiple sources of heterogeneity such as alternative conformations of proteins, partial complexes or unfolded species which will equally be quantified. Thus, incubation conditions such as the buffer used, pH, salt concentrations, sample purity and viability (i.e. folded or unfolded) must be controlled for, to ensure that measurements are representative of the intended system state.

The next step involves digesting cross-linked proteins to the peptide level in order to later identify them via proteomics. Since XL-MS protocols operate on the peptide level, there is no limit on the size and complexity of the system that can be studied[45]. Digestion typically utilises trypsin which is a highly specific and aggressive protease that preferentially cleaves at the C-terminus of Arg and Lys[51]. However, the majority of protein cross-linking, involves linkages between Lys and thus the effectiveness of tryptic digestion is hindered by the presence of cross-linkers[52]. Combinations of secondary enzymes able to cleave at alternative sites such as N-terminal of Asp have been used in conjunction with trypsin to improve the yield of observable peptides[53].

Samples can be subjected to digestion via two commonly employed methods: in-gel digestion, which first subjects cross-linked samples to sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) prior to digestion[54] (Figure 1.10c), or in-solution digestion, in which trypsin and/or other enzymes are added directly to the cross-linked sample. In-gel digestion is particularly complementary to studies of protein complexes in which unwanted partial complexes or oligomeric states may be present alongside the complex of interest. Separation of these species via SDS-PAGE allows bands corresponding to the wanted complex to be isolated through physically excising the band followed by treatment with trypsin. However, the recovery of cross-linked peptides following in-gel digestion is typically poor relative to in-solution methods[55].

The digestion of cross-linked proteins leads to the generation of four categories of peptides based on their cross-linking features (Figure 1.10d). The nomenclature of these categories follows suggestions by Schilling *et al*. and are: 1) non-cross-linked peptides; 2) Type 0 or mono-links where only one reactive group has attached to a peptide; 3) Type 1 intra-peptide cross-links and 4) Type 2 inter-peptide cross-links[56]. Each peptide type reveals structural information regarding the state of the initial captured protein[57]. For example, Type 0 cross-links are akin to covalent labels which provide information regarding regional accessibility, but also the availability of second cross-link acceptor sites in the immediate vicinity of the initial tether point.

In the next step, the identity of each peptide in the mixture must be determined. In XL-MS experiments, cross-linked peptides constitute a minority of the total number of peptides in the mixture and thus a number of procedures have been developed to 'enrich' for cross-linked peptides, improving their observation[52]. A ubiquitously applied method in most XL-MS workflows, involves first subjecting the peptide mixture to LC for separation via differences in their retention time, leading to higher concentrations of cross-linked peptides being detected[53]. Separated peptides enter the MS and are next subject to fragmentation via MS/MS. Peptide identification through MS/MS forms the basis of proteomic techniques and involves the acquisition of both the mass and fragmentation spectra of each peptide (Figure 1.10e). The mass of each intact peptide and its resulting fragment ions are input into a database search to identify and score the peptide against a list of possible candidates from a known sequence[57]. The spectra acquired from cross-linked peptides are more complex than those encountered from standard proteomics experiments due to the presence of the cross-linker and any attached peptides. The exact site of cross-linking can be determined when fragment ions upstream and downstream of the cross-link site can be identified[57]. Fragmentation of peptides can be performed through a number of different methods which result in specific peptide fragmentation patterns (Figure 1.12)[57]. The ability to predict the expected fragment ions based on the fragmentation method used, aids in identification of the

peptide since only a subset of potential fragments can be generated. Collision-induced dissociation (CID) involves accelerating ions into helium, argon or diatomic nitrogen, leading to kinetic fragmentation at the peptide bond[58], generating an N-terminal and C-terminal fragment, referred to as b and y-ions respectively. Other methods such as electron-transfer dissociation (ETD) utilises chemical reagents which transfer electrons to peptide ions, causing fragmentation at N-C_α bonds[59] and leads to generation of c and z-ions[57].

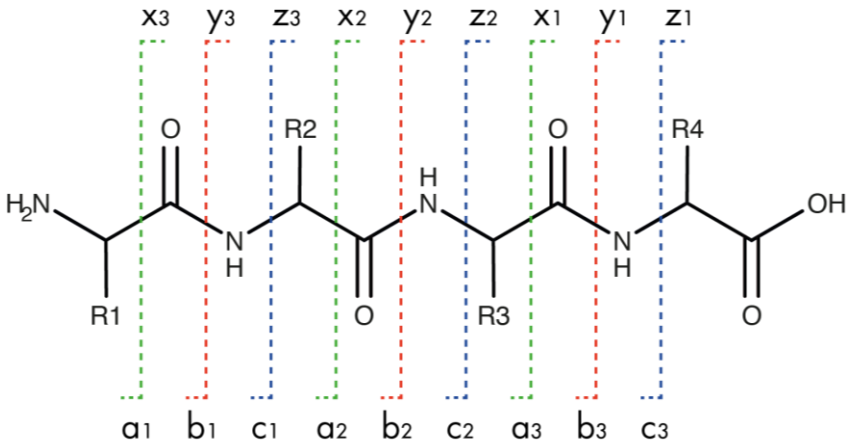


Figure 1.12. Nomenclature of fragment ions produced from fragmentation of the peptide backbone. Fragmentation of peptides at different bonds generates pairs of fragments. N-terminal fragments are known as a, b or c-ions and C-terminal fragments as x, y and z-ions respectively. The number of residues within the fragment is represented by the subscript.

Cross-links identified from XL-MS are exported in tabulated formats by database software such as pLink (Figure 1.10f)[60]. Data tables exported from pLink includes a breakdown of all cross-links identified between a set of supplied sequences which reflect known proteins within the initial cross-linked sample. This breakdown includes which proteins the cross-link was observed between, the exact sequence and residue numbers, the number of identified spectra, observed peptide charge states and precursor mass, as well as a confidence score of the peptide assignment. The resulting data tables from XL-MS experiments can be visualised through a number of different methods. One of these is the commonly used circular plot format produced by the XVis webserver[61], in which residue-scale bars corresponding

to proteins are arranged in a circular ring and each connection represents a cross-link between two proteins at specific residues (Figure 1.10g, right). Two-dimensional representations of cross-linking data are particularly useful when considering proteins without solved structures. If structures or models are available, cross-links can be directly projected onto these models using molecular graphics software such as visual molecular dynamics (VMD)[62] or PyMOL[63] (Figure 1.10g, left).

Cross-links determined from a dynamic protein, contains information regarding its conformational space. Cross-links from multi-subunit protein complexes, capture the relative orientations and positions of individual proteins, including any variations in their dynamics. The utility of XL-MS in studies of protein structure and function, rely on its integration with computational methods that aim to build accurate models through using the observed cross-links to guide their generation. This restraint-based modelling approach is not only limited to cross-links from XL-MS but can also take advantage of any complementary data that can be modelled, including models from X-ray crystallography or density maps from cryo-electron microscopy. The use of cross-linking data for protein complex modelling will be demonstrated in Chapter 4 of this thesis.

1.10 Hydrogen-deuterium exchange mass spectrometry

1.10.1 Principles of hydrogen-deuterium exchange

Hydrogen deuterium exchange (HDX)-MS is a structural MS technique which has garnered attention for its ability to assess protein-protein and protein-ligand interactions, protein folding, and the associated dynamics of these processes, all while within a suitable "native" solution environment[64,65]. The labile amide hydrogens of the protein backbone undergoes exchange with the surrounding deuterium, forming the basis of HDX. Whether or not HDX occurs, depends on both structural and physiochemical factors. The intimate relationship between protein

structure and dynamics thus means that HDX is sensitive to changes in both the former and latter. Exchange of hydrogen for deuterium ("on/in-exchange"), leads to measurable increases in mass, that can be detected via MS. The combination of HDX with bottom-up proteolytic fragmentation strategies, alleviates restrictions on the upper limits of protein size, commonly suffered by other techniques favoured for the dynamic characterisation such as NMR. Further combination of MS allows the localisation of HDX to the peptide-level, providing a read out of the conformational state of the system.

HDX is an isotopic labelling method that involves the exchange of hydrogens in a protein of interest, with deuterium within a deuterium-rich labelling buffer. Hydrogens in N-H, O-H and S-H groups will naturally exchange with the deuterium through acid or base catalysis. However, differences in the exchange rates of each of these chemical groups means that only N-H exchange events are captured by MS. Thus, HDX measurements are concerned with only the exchange of amide hydrogens of the protein backbone.

All residues with the exception of proline and the N-terminal residue feature backbone amide hydrogens that can undergo HDX. The probability of exchange is dependent on structural factors including solvent accessibility and hydrogen bonding status – both of which are intimately related to the secondary structure of the protein. Backbone amide groups are directly involved in the majority of α-helical and β-sheet folds and thus hydrogens of these structures are less likely to exchange (Figure 1.13). Conversely, the fastest exchange occurs in non-hydrogen bonded solvent accessible regions such as those found in flexible or disordered loops.

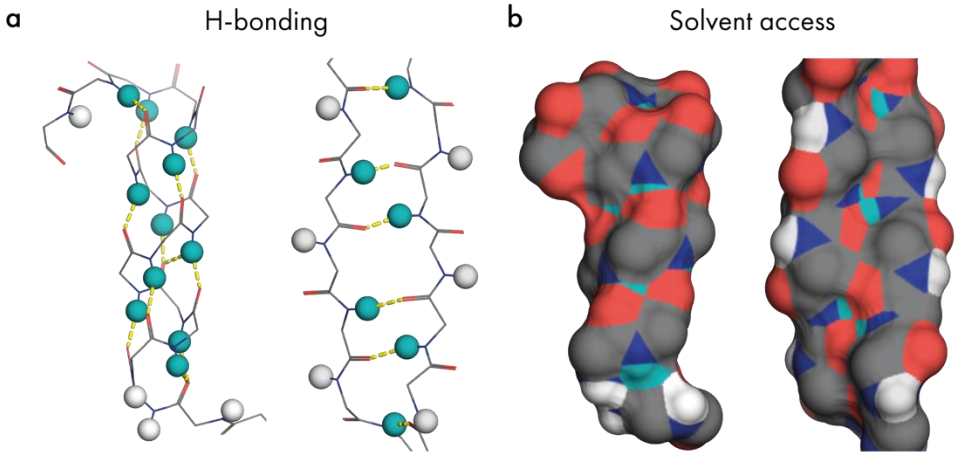


Figure 1.13. Hydrogen bonding and solvent accessibility of α-helical and β-sheet structures. (a) Representative α-helix and β-hairpin of ubiquitin (1UBQ). Backbone shown as lines, amide hydrogens as spheres. Teal and white indicate hydrogen bonded (yellow dashes) and non-bonded hydrogens respectively. Red, blue and grey are oxygen, nitrogen and carbon respectively. Side chains hidden for clarity. (b) Surface view of (a) showing surface accessibility of each atom (PyMOL with solvent probe of 1.4Å). The abundance of colour for each atom are indicative of its solvent accessibility.

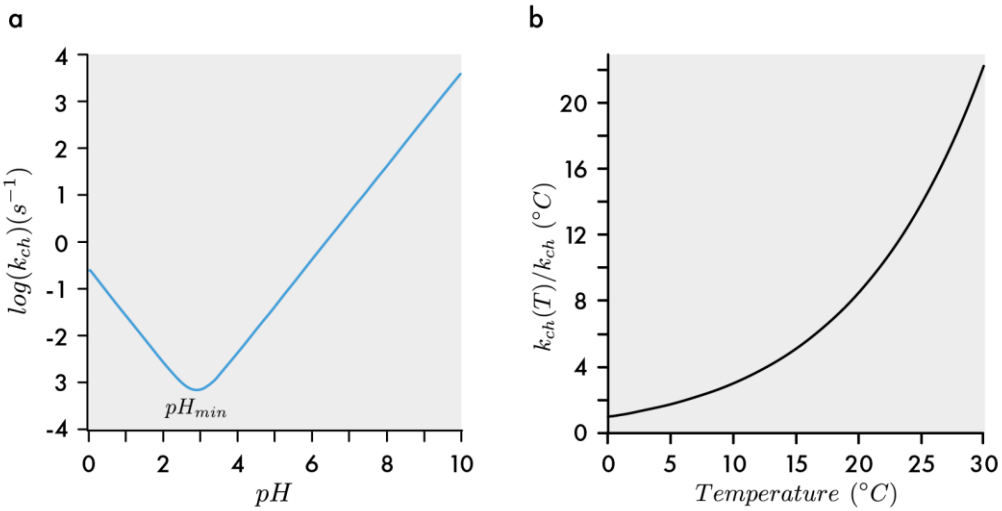


Figure 1.14. Effect of pH and temperature on HDX chemical exchange rate. Chemical rate constant $\boldsymbol{k_{ch}}$ as a function of (a) pH and (b) temperature. The pH at which $\boldsymbol{k_{ch}}$ is at a minimum is marked by $\boldsymbol{pH_{min}}$. Adapted from Bai *et al*. 1993 [66].

The rate of exchange of an amide when unhindered by accessibility or hydrogen bonding is represented by the chemical rate constant[m] k_{ch}, which is in turn affected by a number of physiochemical factors such as the chemistry of the flanking side chains, temperature and pH[65]. The side chains of residues adjacent to amide groups affect k_{ch} either through sterically reducing solvent accessibility (in the case of large non-polar side chains) or through the inductive effect which facilitates exchange by promoting nucleophilic attack of the amide hydrogen in base catalysis.

Using a poly-DL-alanine model peptide, Bai *et al*. demonstrated that k_{ch} is dependent on both pH and temperature[66]. Plotting k_{ch} as a function of pH generates a characteristic V-shaped curve that has a k_{ch} minima at a pH of approximately 2.5 (pH_{min}) (Figure 1.14a). At pH_{min}, the rates of acid and base catalysis are equal, thus when pH is greater than pH_{min}, exchange via base catalysis is dominant. Likewise, when pH is lower than pH_{min}, exchange occurs via acid catalysis[64]. The half-life of exchange for poly-DL-alanine at physiological pH, equates to approximately 1 second, while at pH_{min}, increases to 25 minutes[64]. Similarly, k_{ch} rapidly increases as a function of temperature and is the lowest at 0 °C (Figure 1.14b).

All proteins within their native environments exhibit conformational and thermal fluctuations which lead to transient changes in their structure, commonly thought of as their "breathing motions" [65]. As part of these motions, individual regions or the entire protein may temporarily become more solvent accessible or hydrogen bonds may be transiently broken, allowing HDX to take place. The rate of opening and closing transitions are each associated with the rate constants k_{op} and k_{cl} and thus HDX can be described using the following mechanism (1.12):

$$N - H_{closed} \underset{k_{cl}}{\overset{k_{op}}{\rightleftharpoons}} N - H_{open} \xrightarrow[D_2O]{k_{ch}} N - D_{open} \underset{k_{op}}{\overset{k_{cl}}{\rightleftharpoons}} N - D_{closed} \qquad (1.12)$$

m Also referred to as the intrinsic rate constant

Under this mechanism, two exchange regimes have been observed. The first, EX1, occurs when k_{ch} is faster than k_{cl}, that is, the opening of amides is long enough such that exchange can occur (1.13)[64]. In EX1, isotopic exchange rate k_{HDX} is therefore equal to k_{op}. The opposite is observed in EX2 kinetics whereby multiple rounds of opening and closing occur before exchange takes place (1.14). Under EX2, k_{HDX} is equal to the ratio between open and close rates represented by the equilibrium constant K_{op} and k_{ch}.

$$EX1: \quad k_{ch} \gg k_{cl}$$
$$k_{HDX} = k_{op} \tag{1.13}$$

$$EX2: \quad k_{cl} \gg k_{ch}$$
$$k_{HDX} = \frac{k_{op}}{k_{cl}} k_{ch} = K_{op} k_{ch} \tag{1.14}$$

Under physiological conditions, the majority of exchange occurs under the EX2 regime[65]. The opening and closing motions exhibited by a protein while within a native environment, is a feature of its conformational dynamics. Each EX2 exchange event relates to a particular conformation of its solution ensemble that permits exchange to take place at the particular amide hydrogen. Therefore, understanding the relationship between the HDX and the underlying conformational dynamics of proteins, is paramount to making full use of the structural information available from HDX-MS.

1.10.2 The HDX-MS workflow

Advances in the understanding of HDX and its utility as a conformational reporter of proteins has led to the emergence of several different HDX-MS methods, such as those that can derive binding affinities (e.g. PLIMSTEX[n]) and monitor protein

n PLIMSTEX: Protein-ligand interactions by mass spectrometry, titration and HDX

unfolding events (e.g. SUPREX[o]). PLIMSTEX experiments incorporate ligand titration methods into the HDX workflow in order to generate binding curves from which affinity values can be mathematically extracted[67]. SUPREX on the other hand performs HDX in the presence of varying concentrations of denaturants and so captures conformational information along the protein's unfolding pathway[67]. The most commonly applied HDX-MS routine, however, remains the use of HDX in time-resolved experiments for monitoring binding events and changes in protein conformation. The following section will outline various key steps of the HDX-MS workflow.

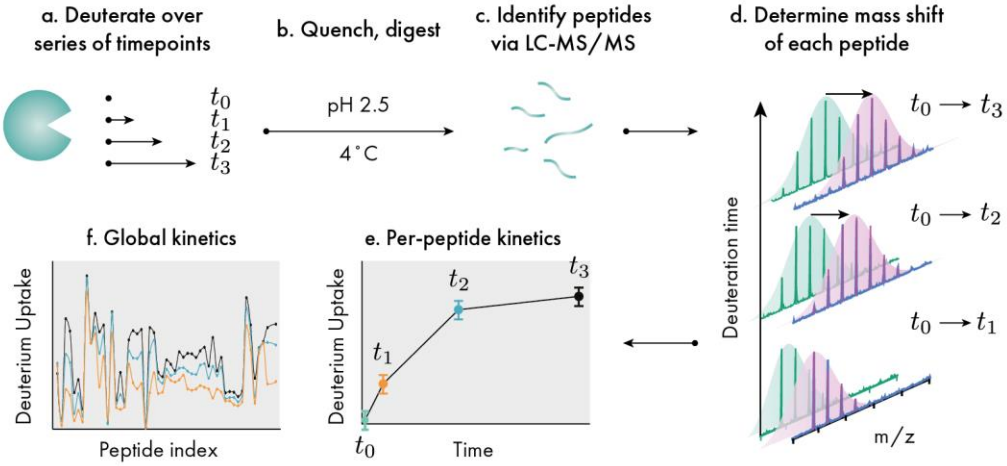


Figure 1.15. Workflow of HDX-MS kinetics experiments. (a) A protein of interest is incubated in deuterium-containing buffer for a series of timepoints. A non-deuterated t0 timepoint is also processed to provide a reference mass. (b) Exchange is quenched by increasing the acidity of the mixture and dropping its temperature to 0˚C. (c) The deuterated protein is digested to the peptide level and input into an LC-MS/MS workflow to acquire *m/z* and fragmentation spectra for each peptide. The identity of each t0 peptide is determined through a database search. (d) The mass of each peptide at different t is compared to t0 to determine the mass shift as a result of deuterium uptake. (e) Deuterium uptake as a function of time can be displayed for each peptide in kinetics plots. (f) Deuterium uptake for all peptides can be displayed via a number of different global plots.

o SUPREX: Stability of unpurified proteins from rates of HDX

In time-resolved HDX-MS experiments, a protein of interest is exposed to a deuterium-rich buffer at physiological pH, for a series of deuteration timepoints (Figure 1.15a). A non-deuterated sample is set up in parallel and subjected to the same workflow to serve as a reference state representing deuteration at $t = 0$. Under physiological pH, the half-life of exchange as demonstrated earlier by Bai *et al*. is approximately on the order of seconds[66]. The timepoints selected are typically spaced out such that they sample the level of deuteration from seconds to hours or even days. Long timepoints serve the utility of representing the fully deuterated state and can also be used for back-exchange correction, explained later n this thesis. The effect of temperature and pH on the intrinsic exchange rate of amides can be taken advantage of in order to halt exchange. To prevent further exchange, a quench buffer is injected into the labelling solution in order to rapidly drop the mixture to a pH of 2 and 4 ° C (Figure 1.15b). Under these conditions, the half-life of exchange increases to between approximately tens of minutes to hours[68]. While exchange is slowed under quench conditions, a degree of "back-exchange" (in which N-D is exchanged with N-H) can occur, leading to a loss of the deuteration signal[69]. The degree of back-exchange is both dependent on the residue and on the post-labelling experimental set up. Quench conditions also typically include the addition of chaotropic agents such as urea or guanidine hydrochloride and reducing agents such as dithiothreitol (DTT) which promotes protein unfolding and reduction of any disulphide bonds which may prevent efficient digestion in the next step.

Following quench, the protein mixture is subjected to an LC method which performs digestion to the peptide level and separation prior to ESI (Figure 1.15c). Digestion involves injection of the quenched sample into a chromatographic column containing a matrix of immobilised proteases. To minimize back exchange, it is necessary to maintain quench conditions during digestion and thus the protease employed must be active under these conditions. Another criteria of the protease is that its digestion must be non-specific since the spatial resolution of HDX relies on overlapping peptides. A commonly used enzyme for this purpose is pepsin, however

greater enzymatic efficiency has been observed from alternative enzymes including nepenthesin from the *Nepenthes* genus[70] and the use of multiple chained enzymatic columns has also been explored[71]. Elution from the enzymatic column proceeds with injection into a trapping column and reverse phase UPLC for peptide separation, followed by ESI into the MS. Within the mass spectrometer, peptides are subjected to MS/MS in a similar manner as that described for the XL-MS workflow. The *m/z* and fragmentation spectra for each peptide is collected and can be analysed in the next stage.

1.10.3 Data processing and experimental types

Data processing methods following spectral acquisition depend on the instrumentation used, and a number of different software have been developed for the purpose of streamlining HDX-MS data processing. For the Waters Synapt line of HDX-MS instruments, data processing is performed in two stages: first, a proteomics database search using the ProteinLynx Global Server (PLGS) which identifies and generates a list of observed peptides based on *m/z* and fragmentation spectra; second, peptide mass assignment using DynamX which provides facilities for automated and manual evaluation of spectra. The role of data processing software such as PLGS and DynamX is to allow extraction of peptide mass information following deuteration as a function of time (Figure 1.15d). The mass of each deuterated peptide is compared with the mass of the non-deuterated reference in order to determine the mass shift associated with deuteration. HDX-MS data processing software will be covered in much greater detail in Chapter 3: of this thesis.

Chapter 2: Developing a workflow for modelling protein flexibility using ion-mobility MS and gas phase simulations

Preface

The following publication has been included in this thesis chapter in the '*Thesis Incorporating Publications*' format:

Hansen, K.[†], Lau, A. M.[†], Giles, K., McDonnell, J. M., Struwe, W. B., Sutton, B. J., Politis, A. (2018). A mass spectrometry-based modelling workflow for accurate prediction of IgG antibody conformations in the gas phase. *Angewandte Chemie International Edition*, 57 (52): 17194–17199.

[†] Denotes authors of equal contribution.

Author contributions

As co-first author of the Hansen and Lau *et al*. 2018 publication, I designed and performed the computational modelling and gas phase simulations presented in the paper along with Kjetil Hansen. Kjetil Hansen and Dr Weston Struwe are responsible for all experimental lab work and biological interpretation. I created all figures, and both myself and Kjetil Hansen led the writing of the manuscript and peer review process. Professor Kevin Giles, Professor James McDonnell, Professor Brian Sutton and Dr Argyris Politis played supervisory roles for this project.

A mass-spectrometry-based modelling workflow for accurate prediction of IgG antibody conformations in the gas phase

2.1 Abstract

Immunoglobulins are biomolecules involved in defence against foreign substances. Flexibility is key to their functional properties in relation to antigen binding and receptor interactions. Here we develop an integrative strategy combining ion mobility mass spectrometry (IM-MS) with molecular modelling to study the conformational dynamics of human IgG antibodies. Predictive models of all four human IgG subclasses were assembled and their dynamics sampled in the transition from extended to collapsed state during IM-MS. Our data imply that this collapse of IgG antibodies is related to their intrinsic structural features, including Fab arm flexibility, collapse towards the Fc region, and the length of their hinge regions. The workflow presented here provides for the first time an accurate structural representation in good agreement with the observed collision cross section for these flexible IgG molecules. These results have implications for studying other non-globular flexible proteins.

2.2 Introduction

Immunoglobulins (Ig), or antibodies, are the proteins responsible for mediating an extensive network of immunological responses. The past decades have seen a steady increase of interest in developing Igs as biotherapeutic agents for the treatment of various diseases, including cancer and autoimmune disorders[72-74]. While the architectures of Igs are relatively conserved, they exhibit dramatic differences in their dynamics and mode of interactions with antigens and cognate receptors[75,76]. These differences stem from intrinsic features in their structures such as binding site specificity and hinge flexibility (Figure 2.1a)[77].

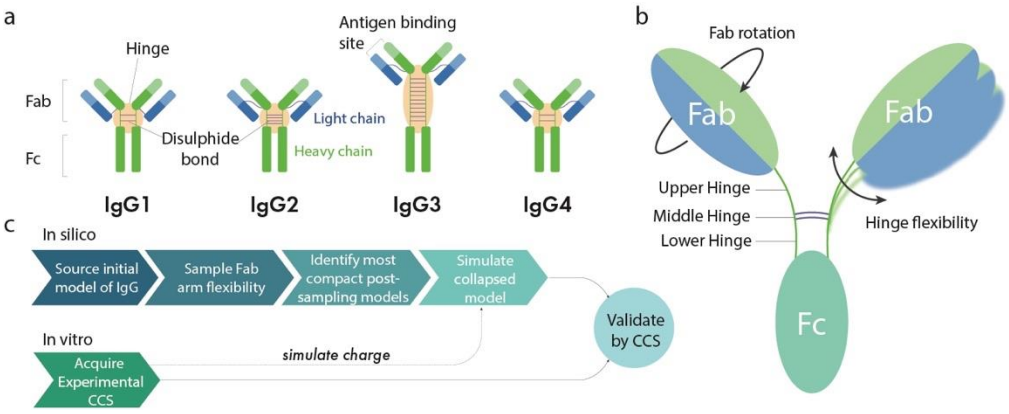


Figure 2.1. Schematics and workflow for modelling antibody flexibility. (a) Schematic representation of human IgG1-4 subclasses. (b) Representative structure of IgG1, denoting hinge substructure and modes of Fab movement stemming from the upper hinge. (c) Integrative workflow generating and comparing the calculated *CCS* values of initial, post-sampling and gas phase MD models with experimental *CCS* values.

There are five isotypes or classes of Igs, the most abundant of which in humans is Ig gamma (IgG), comprising approximately 75% of all human antibodies in serum[78]. IgG is by far the most commonly exploited isotype for biotherapeutics[73], including, bispecific antibodies[79,80] and antibody-drug conjugates (ADCs)[81-83]. In 2017, 10 new antibody therapeutics were approved, all of which were IgG based[84]. There are four subclasses of human IgG, named IgG1-4 (Figure 2.1a). While IgG1, 2 and 4 are similar in topology, overall length and hinge length, IgG3 has a markedly longer hinge,

producing a molecule much longer than the other subclasses[85,86]. IgG molecules exhibit a high degree of heterogeneity due to their extensive glycosylation, and also sequence variability in their antigen binding regions. All IgG molecules consist of two heavy chains and two light chains that are covalently linked via disulphide bridges in a characteristic "Y" shaped topology (Figure 2.1b). A central hinge separates two Fab "arms" from the Fc "stem" of the IgG molecule. This hinge plays a pivotal role in providing IgG molecules with flexibility, allowing relative Fab-Fab and Fab-Fc movements[87]. The hinge and Fc region play an important role in binding immune effector proteins including, the Fc gamma receptors (FcγR), neonatal Fc receptor (FcRn) and complement component C1q[86] (Supplementary Figure 6.1). The ability for all IgG subclasses except IgG4 to trigger the complement cascade via C1q[88], for example, illustrates that the intrinsic structure and dynamics of these molecules have functional consequences for each of the IgG subclasses.

Native mass spectrometry (MS) has recently emerged as a powerful method for interrogating proteins and their complexes, providing valuable information about their stoichiometry and topology[9,89-95]. Native MS can be hyphenated with IM; the resulting ion mobility (IM)-MS method offers an extra dimension enabling shape information on the investigated proteins. IM-MS allows for derivation of topological information of proteins through calculating their collisional cross section (*CCS*). *CCS* is described as the rotationally-averaged cross section of a molecule and is calculated based on the overall size and molecular architecture[15]. The experimentally measured *CCS* can be compared to theoretical *CCS* calculated from structural models derived by molecular dynamics (MD) simulations or other modelling techniques[23,96], enabling structures to be assigned back to experimental observations[97].

Native MS mainly uses electrospray ionisation (ESI) for the purpose of creating multiply charged protein ions[98]. The response of folded proteins entering the gas phase through ESI is most commonly described through the charged residue model (CRM)[2,5,99]. The CRM envisions gradual droplet desolvation leading to production of

a dry protein ion. While the behaviour of a globular protein transferring into the gas phase of a mass spectrometer can be rationalised under the CRM framework, here we pose the following question: do these same rules apply to non-globular and flexible proteins? Early studies which compared the experimental *CCS* of antibodies to those calculated from their crystal structures, observed a >30% discrepancy between these *CCS* values[28,29], suggesting collapse of the protein in the gas phase. Such collapse is experienced by non-globular molecules that are intrinsically flexible or disordered in solution and are capable of conformational change[21,30,100,101]. Whilst others have explored simulating such collapsing structures[100], these call for computationally complicated methods such as "trajectory stitching"[42] or including mobile proton algorithms[44,102], which may be impractical for large molecules. Here, we have developed an integrative IM-MS-based strategy which enables the prediction of the structure and dynamics of IgG molecules in the gas phase including, for the first time, capturing and simulating the dynamics of human IgG3 (Figure 2.1a-c). In the first step, homology models of the antibodies were built and subsequently subjected to Fab arm sampling allowing representation of their intrinsic flexibility as a conformational ensemble. Simultaneously, we subjected IgG1-4 to IM-MS experiments and derived their corresponding *CCS* values. The most compact conformations were taken forward for vacuum MD simulations in order to model the gas phase structure of IgG molecules.

2.3 Materials and Methods

2.3.1 Sample Preparation

IgG1, IgG2, IgG3 and IgG4 (kappa from human myeloma plasma) were purchased from Sigma-Aldrich at a concentration of 1 mg mL^{-1}. Lyophilised intact mAb mass check standard was purchased from Waters. Herceptin was purchased from the Churchill Hospital Pharmacy, University of Oxford. The proteins were buffer exchanged into 150 mM ammonium acetate pH 7 using Micro Bio-Spin columns (Bio-Rad) prior to running on the mass spectrometer. Transthyretin (TTR; 56 kDa), alcohol dehydrogenase (ADH; 148 kDa) and glutamate dehydrogenase (GDH; 336 kDa) were used as *CCS* calibrants and were buffer exchanged using the same procedure as above. The glycans were removed for the deglycosylation experiments with PNGase F (New England BioLabs) for 4 hours at 37 °C.

2.3.2 Ion Mobility

A commercial qTOF-TWIMS instrument (Synapt G2-Si, Waters) was used with a nano-ESI source. The instrument was run with positive polarity in sensitivity mode and calibrated with caesium iodide. Capillaries were pulled in-house with a Flaming/Brown P-97 micropipette puller (Sutter Instruments) and coated with Au:Pd (80:20) using a sputter coater (Quorum Q150RS). The following mass spectrometer settings were used: capillary voltage 1.3-1.8 kV, sample cone 50 V, source temperature 45 °C, trap pressure $3.6x10^{-2}$ mbar, drift tube pressure 2.6 mbar, TOF pressure $1.0x10^{-6}$ mbar, IM-MS wave height 40 V, *m/z* range 500-12000, cone gas 0 L hr^{-1}. Nitrogen was used as the ion mobility gas and drift times were collected at IM-MS wave velocities of 550, 600 and 640 ms^{-1}.

The resulting data was processed using MassLynx V4.1 (Waters Corp. Manchester, UK) and PULSAR[18] which contains literature values for the *CCS* calibrants[103]. CCS_{exp} of

TTR, ADH and GDH were used to generate calibration curves (for each T-wave velocity) to which the IgG1-4 data points were fitted (R^2 = 0.985-0.989). Final CCS_{exp} for each IgG1-4 were taken from the lowest charge state species. All CCS_{exp} values were converted to CCS_{He} in PULSAR.

2.3.3 High-Resolution Native Mass Spectrometry

A Thermo Q-Exactive mass spectrometer (Thermo Fisher Scientific) modified for detection of high molecular weight ions was used for IgG1 glycosylation analysis. Data was obtained in positive ion mode with an acquisition window of 1000 to 15000 *m/z*. Ions were desolvated in the HCD cell with 100 V. Additional settings were as follows: capillary voltage = 0.8-1.0 kV; source temperature = 60°C; max injection time = 100 ms; S-lens RF = 150; resolution = 17500. Spectra were obtained with 10 microscans, averaged over 50 scans. Data was processed using XCalibur 2.1 software (Thermo Fisher Scientific, Germany) and glycoforms were assigned manually.

2.3.4 Generating initial models of IgG1, IgG2 and IgG4

All homology modelling was performed using MODELLER[104]. IgG1 (UniProt accession: P01834 and P01857) was modelled using PDBs 1HZH (human) and 1IGY (mouse) as template. The Fab of 1HZH (chains B and D) missing covalent connection to the rest of the molecule, was extracted and aligned to the Fab of 1IGY in order to recover an extended solution-like conformation of IgG1. The structure of IgG2 (AN: P01834 and P01859) was modelled using PDB 1IGT (mouse; whole molecule), 4L4J (human; Fc) and 2QSC (human; Fabs). Two additional disulphides were inserted into the hinge to generate a representative model of human IgG2. 200 models of each IgG1 and IgG2 were generated and evaluated based on their discrete optimised protein energy (DOPE) score[105]. Missing residues of the human IgG4 crystal structure were re-generated automatically in MODELLER using PDB 5DK3 (human) as template. Glycans structures were not modelled into any of the IgG molecules for two reasons. Firstly, each IgG exhibits numerous glycoforms which dramatically

increases the number of starting models of our study, both for the Fab arm sampling and gas phase simulation sections. Secondly, deglycosylation of IgG molecules results in no significant difference in experimental *CCS*. These observations have led us to believe that the added complexity of including glycan structures does not offer significant benefits to our modelling workflow.

2.3.5 Homology modelling of IgG3

For IgG3 which lacks most structural representation, we acquired fragments of the structure from PDBs 4HAF (human; Fc) and, 4HDI and 1CLZ (mouse; Fab) and manually built the hinge structure using 11 CYS-CYS pairs interspersed with six tri-proline helices. All other hinge residues were automatically added with MODELLER. Glycans were not modelled for IgG3, as described in the homology modelling procedure of IgG1, 2 and 4. The IgG3 model was then subjected to 100 ns of explicit solvent molecular dynamics simulation in GROMACS 5.1.3[106] with the CHARMM27 (modified CHARMM22 for proteins) forcefield[107]. The atomistic homology model of IgG3 was added to a triclinic simulation box (178 x 169 x 252 Å) with an edge buffer of 10 Å to account for flexibility and prevent interactions with periodic images. Disulphide bonds were manually checked to ensure correct bonding. 243,611 TIP3 waters and 2 chloride counterions were added to neutralise the system charge. We then performed energy minimisation using a steepest-descent algorithm, followed by equilibration in isochoric-isothermal (300 K, τ = 0.1 ps) and isobaric-isothermal (1.0 bar, τ = 2.0 ps) ensembles for 1 ns each. Equilibration employed the "V-rescale" modified Berendsen thermostat and Parrinello-Rahman barostats. The LINCS algorithm was employed to restrain bonds. Finally, production simulation of the system was continued for 100 ns at constant temperature and pressure. For non-covalent interactions, we utilised particle mesh Ewald (PME) with a grid spacing of 0.16 nm for long-range electrostatic interactions, and the Verlet cut-off scheme for Van der Waal calculations. The RMSD evolution of the simulation was monitored and reviewed after 100 ns of simulation to ensure appropriate convergence of the IgG3

structure. To extract a single representative model of IgG3, we clustered models from the final 50 ns of the simulation and identified centroid model of the major conformation.

2.3.6 Fab arm conformational sampling

For conformational sampling of each of the IgG1-4 Fab arms, we first identified the selection of residues which constituted their upper hinges. For each IgG heavy chain, these were, IgG1: D446-T450, IgG2: E437-K439, IgG3: E219-T230, IgG4: E437-P443. The conformational space of each upper hinge and its Fab were then sampled using a rapidly exploring random tree (RRT) algorithm available from the Integrative Modelling Platform (IMP)[108]. This procedure sets the disulphide top-most disulphide of each hinge as the tree root and randomly tests availability for each node (connected atom) to rotate to a new position which is also permissible by the residue's torsional space. Conformations which do not result in steric clashing or overlap are exported as a structure within the ensemble. Both Fab arms are sampled simultaneously with 10,000 models being generated in total for each IgG1-4.

2.3.7 Gas phase molecular dynamics simulations of IgG1-4

Each of the lowest *CCS* conformations of IgG1-4, including two models selected from the pool of lowest 50 *CCS* models, were subjected to gas phase molecular dynamics simulations using GROMACS 5.1.3[106]. Since there is no method of determining the experimental charge sites, we pre-charged our IgG models using a localised charge model. This model reflects the lowest observed experimental charge state (21+ for IgG1, IgG2 and IgG4, 22+ for IgG3). Charges were applied to a randomly selected distribution of basic (lysine, histidine and arginine) residues which were found within 5 Å residue depth of the protein surface, using a combination of the DEPTH server[109] and in-house scripts. Pre-charging was repeated where charges were placed too close to each other or prevented the structure from collapsing. We also did not consider acidic residues or neutral salt bridges due to there being no method of

accounting for these interactions experimentally. All acidic residues (aspartate and glutamate) remained neutral for our gas phase simulations. Simulations were performed using the OPLS forcefield due to the availability of protonated arginine topologies. All disulphide bonds were manually checked to ensure correct bonding. Energy minimisation was performed for 50,000 iterations using a steepest descent minimiser, followed by position restraints for all bonds for 500 ps. Simulations were carried out at a temperature of 300 K and regulated using the Berendsen thermostat (τ = 0.1 ps). Pressure coupling and periodic boundary conditions were switched off due to the *in vacuo* nature of the simulations. A cut-off scheme of infinite distance was used for coulombic and van der Waals interactions. Each model was equilibrated briefly at the correct temperature for 1 ns and then for a further 10 ns to produce the collapsed topologies. RMSD, radius of gyration and *CCS* was monitored throughout all simulations.

2.3.8 Gas phase simulations of IgG4 for charge states 22-25+

All simulations were carried out as detailed above. Each simulation begins from an identical pre-collapsed model of IgG4 (produced by Fab arm sampling). 22, 23, 24 and 25 charge sites were selected randomly using in house scripts. The charge site distributions are different between each simulation. Each simulation was performed for a total of 10 ns, and the average *CCS* and *CCS* variation over the last 1 ns of simulation time was calculated.

2.3.9 *CCS* Calculation of Computational Models

All *CCS* for IgG structures were calculated as CCS_{He} using IMPACT software[23]. *CCS* were calculated through scaling the projection approximation (PA) from IMPACT, by a factor of 1.14 to account for PA underestimation. PA measurements from IMPACT have been calibrated to values calculated from the trajectory method, with a root mean square relative error of less than 1%[23]. This linear scaling factor of PA has shown success with approximating the experimental *CCS* of large protein

complexes[22]. While other direct *CCS* approximation methods such as exact hard sphere scattering (EHSS)[26], trajectory method (TM)[110] and projection superposition approximation (PSA)[111] are available, the magnitude of models generated in our study (minimum of 50,000), required a high throughput calculation method such as IMPACT[23].

2.4 Results

Due to the flexible nature of IgG molecules, there are currently only four intact IgG crystal structures available: IgG1 (human: 1HZH, mouse: 1IGY)[112,113], IgG2 (mouse: 1IGT)[114] and most recently, IgG4 (human: 5DK3)[115]. We modelled human IgG1, 2 and 4 using the available crystal structures, followed by generation of any missing residues. IgG3, however, presented a greater challenge due to its complex hinge structure and lack of crystallographic representation. The absence of an IgG3 intact crystal structure is likely due to its greater flexibility. We thus built a homology model of human IgG3 and subjected it to 100 ns of explicit solvent simulation to model its average solution state conformation (Supplementary Figure 6.3). This average conformation exhibits a hinge length of approximately 70 Å between the Fab and Fc regions, with the fully extended length of the hinge at approximately 110 Å (Supplementary Figure 6.4). This is in agreement with an earlier electron microscopy study by Roux *et al*. who observed a distance of 80 ±23 Å between the Fabs and Fc in solution, and an estimated 100-110 Å for the length of the extended hinge[77], as well as with other hydrodynamic and solution X-ray scattering studies[116,117].

Table 2.1. Experimental and model *CCS* values for IgG1-4

Subclass	IgG1	IgG2	IgG3	IgG4
Theoretical mass (kDa)[a]	150	150	170	150
Experimental mass (Da)[b]	149,328 (±89)	154,297 (±42)	162,123 (±4)	155,758 (±62)
Experimental charge[c]	21+	21+	22+	21+
Overall hinge length[d]	12	12	62	12
No. hinge disulphides	2	4	11	2
Upper hinge residues sampled	5	3	12	7
Initial model *CCS* (Å²)[e]	9532	9747	10958	9512
Fab arm sampling *CCS* (Å²)[f]	8756	8597	9170	8484
Δ*CCS* of sampling (Å²)[g]	1102	929	1329	1080
Collapsed model *CCS* (Å²)[h]				
Model 1	7226 (±176)	7396 (±201)	7284 (±173)	7017 (±204)
Model 2	6988 (±196)	7197 (±184)	7176 (±197)	6766 (±268)
Model 3	7142 (±176)	7309 (±213)	7588 (±202)	6644 (±179)
Experimental *CCS* (Å²)[i]	6827 (±81)	7030 (±113)	7173 (±68)	7024 (±97)
Deglycosylated *CCS* (Å²)[j]	6851 (±61)	7087 (±56)	7202 (±43)	7095 (±51)
Net solution charge[k]	20+	2-	2+	2+

a Approximate mass of glycosylated protein given sequence variability in Fc and Fab regions.

b Experimentally observed glycosylated mass via MS (± standard deviation).

c Lowest observed experimental charge for glycosylated proteins.

d Fc to Fab distances (UniProt: IgG1 P01857, IgG2 P01859, IgG3 P01860, IgG4 P01861).

e $CCS = PA \times 1.14$ calculated via IMPACT for starting models.

f Lowest CCS_{model} generated from Fab arm sampling.

g CCS_{model} range of ensemble from Fab arm sampling.

h CCS_{model} for triplicate models after 10 ns of gas phase simulation (± denotes range over final nanosecond).

i Average *CCS* for lowest charge over T-waves 550, 600 and 640 ms^{-1} (±SD) for glycosylated proteins.

j Average *CCS* for lowest charge over T-waves 550, 600, and 640 ms^{-1} (±SD) for deglycosylated proteins.

k Net solution charge of each IgG molecule as determined via the Protparam webserver.

We next subjected all four of the human IgG subclasses to IM-MS allowing us to quantify their topology through their experimental *CCS* values (Figure 2.2a, Supplementary Figure 6.5-Supplementary Figure 6.6, Table 2.1). Consistent with previous studies of human IgG1-4[28,29], we observed an approximately 30% difference between the experimental *CCS* (CCS_{exp}) and model *CCS* (CCS_{model}; Table 2.1), suggesting significant structural collapse. Despite the much longer length of IgG3 compared to IgG1, 2 and 4, all IgG antibody are equally able to collapse producing a *CCS* of approximately 7000 $Å^2$. We further deglycosylated all antibodies and subjected them to IM-MS (Supplementary Figure 6.7-Supplementary Figure 6.8) – the resulting *CCS* showed no significant difference compared to their glycosylated counterparts (<1%) (Supplementary Figure 6.9). To further exclude the possibility that glycoform heterogeneity influences IgG gas phase conformations, we characterised the glycans bound to two additional monoclonal antibodies, Herceptin and Waters mAb, and compared their glycoforms to Sigma IgG1 (Supplementary Figure 6.10). While we identified different N-linked glycans bound to the IgG molecules, their corresponding *CCS* values were found to be within 1%, indicating no significant changes in their conformations.

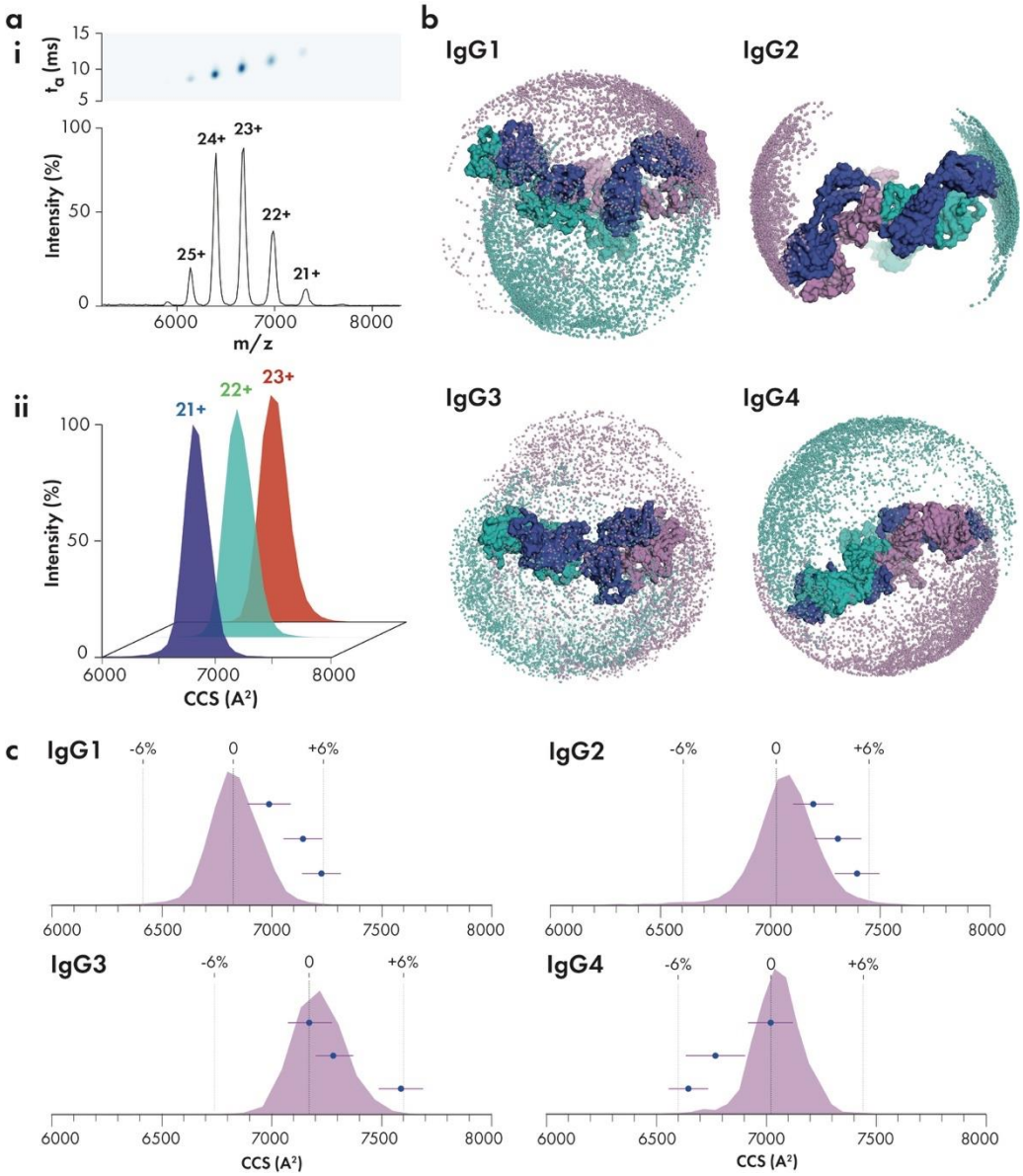


Figure 2.2. Modelling the conformational flexibility of antibodies. (a) Representative mobilogram and native mass spectrum (i) and *CCS* distributions for 21-23+ charge states of IgG2 (ii). (b) Space occupied by IgG1-4 Fabs following upper hinge flexibility sampling. Each sphere represents one model for each IgG Fab heavy chain (teal and purple). Light chains are shown as blue. Initial models are shown as surface representations. (c) Overlay of experimental *CCS* distribution with triplicate simulated collapse models. Experimental error is represented by the ±6% dotted lines. Purple error bars represent the *CCS* range over the last 1 ns of gas phase simulation.

The ability of significant compaction is likely provided by IgG hinges imparting the steric freedom necessary for the Fab and Fc domains to contort into a compact structure. While collapse of IgG structures and other flexible molecules in the gas phase have been widely observed, simulating their collapsed structures remains a

challenge. Previous studies beginning gas phase simulations directly from crystal structures of IgG molecules, show >20% discrepancy from their CCS_{exp} values[28,29]. Thus, we hypothesise that IgG molecules may experience pre-collapse prior to transfer into the gas phase.

With the aim of developing a methodology able to simulate the collapse of IgG molecules, we designed an *in silico* workflow for pairing experimental data with computational models (Figure 2.1c). In the first step of our modelling workflow, we subject each of the four initial IgG models to Fab arm conformational sampling using a rapidly exploring random tree algorithm. This sampling technique produces randomly varied Fab conformations given the degrees of freedom allowed by the antibody's flexible upper hinge residues. Specifically, the conformations of residues between the most N-terminal hinge disulphide bond and the Fab domains are explored. Sampling in this way produces a model ensemble of conformations of IgG structures which mimic their flexibility in a solution environment.

Through Fab arm sampling, we generated 10,000 conformations for each IgG and calculated the model variation over each ensemble (Figure 2.2b, Table 2.1, Supplementary Figure 6.11-Supplementary Figure 6.12). The *CCS* variation in each of the IgG ensembles correlated well with previously reported Fab flexibility (IgG3 > IgG1 > IgG4 > IgG2)[77,118] and upper hinge lengths. Closer inspection of the conformational space occupied by each Fab arm revealed that Fabs of IgG1, 2 and 4 are restricted to their own hemispheres (Figure 2.2b). The Fabs of IgG3 however, share a high degree of overlapping space indicative of their enhanced flexibility provided by the longer upper hinge region (Figure 2.2b).

Our sampling strategy identified conformations of IgG1 which are highly similar to those modelled through solution-based small angle X-ray scattering (SAXS)[119]. A recent study also highlighted the ability of Fab arm sampling and clustering analysis to deliver solution-relevant antibody conformations[120]. Within our IgG2 ensemble, movement of Fab arms are more confined due to the presence of two extra

disulphide bonds located in the upper hinge compared to the other short IgG molecules (IgG1 and IgG4) (Table 2.1). While Fabs of IgG1, IgG3 and IgG4 are able to flex away from their Fc, exposing potential receptor binding sites (Supplementary Figure 6.1), this dynamic behaviour is not shared to such a degree by IgG2. It is interesting to speculate that this modelling observation may offer insight into why experimentally, IgG2 exhibits reduced affinity to some FcγR receptors compared to IgG1, 3 and 4[86,121].

While Fab arm sampling provides a powerful method of exploring the conformational space, it is important to note that this is a pseudo-simulation and does not take the energetic landscape of the molecule into account. To provide structures relevant to the gas phase environment, we subjected models from each ensemble to molecular dynamics (MD) in vacuo (Supplementary Figure 6.13-Supplementary Figure 6.16). We performed simulations in triplicate by selecting the lowest *CCS* and two low *CCS* models from each Fab arm sampling ensemble. Each model was pre-charged with the lowest observed experimental charge state (Table 2.1). The models were then simulated *in vacuo* for 10 ns. All CCS_{model} calculated for the final simulation frame for each IgG were within approximately 6% of the CCS_{exp} values, with the closest match being IgG3 showing 0.1% *CCS* difference. For each set of three simulations, the *CCS* difference between independent simulations were 3.8% for IgG1, 1.9% for IgG2, 4.3% for IgG3 and 4.9% for IgG4. We additionally carried out simulations of IgG4 for charge states 22-25+ which show agreement with experimental values (Supplementary Figure 6.17). To the best of our knowledge, this is the first time that the dynamic structures of substantially collapsed IgG molecules have been modelled with this level of agreement with experimental values.

Overlaying the CCS_{model} of our collapsed models with the experimental *CCS* distribution showed that each individual simulation occupies a narrow *CCS* range, indicating that collapsed structures are relatively inflexible in the gas phase (Figure 2.2c). We hypothesise that IgG flexibility in solution leads to a diverse population of rigid collapsed structures in the gas phase, resulting in the observed experimental

CCS distribution. The results of each step of our modelling workflow has been summarised in Figure 2.3.

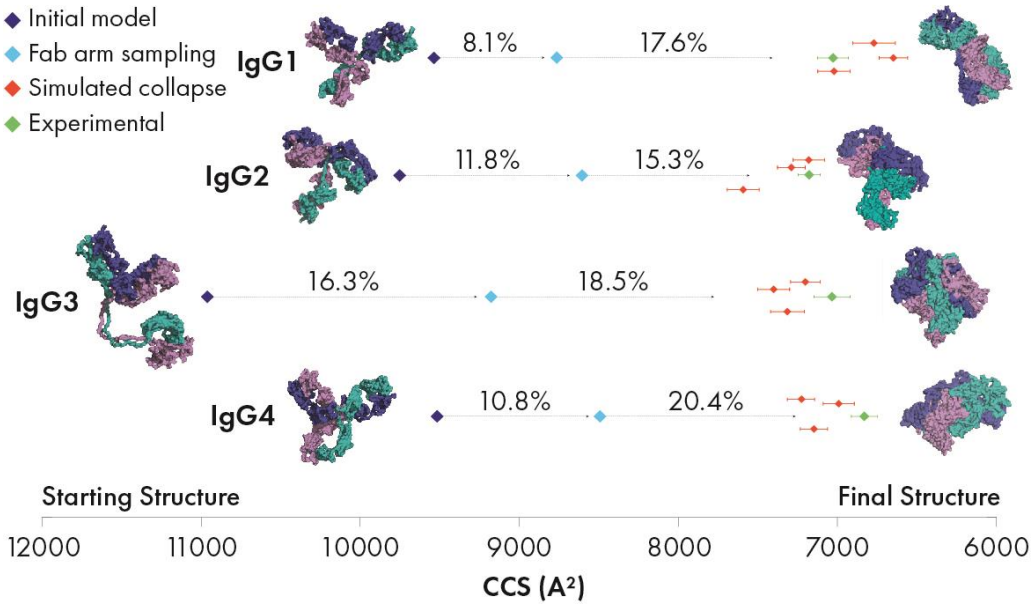


Figure 2.3. Summary of experimental and model *CCS* for IgG1-4. *CCS* was calculated for each stage of the modelling workflow. The reduction in *CCS* between modelling stages is shown by percentages. Error bars for the experimental data points (green) represent the standard deviation of measurement. Error bars for simulated collapse models (red) show *CCS* range over the last 1 ns of simulation.

2.5 Discussion and Conclusions

Overall, our modelling workflow has generated collapsed models of IgG1-4 which closely match experimental *CCS* values. The steps undertaken aim to simulate the collapse of these molecules in the gas phase whilst staying in line with the current consensus of the CRM of folded proteins undergoing ESI. The timeline emerging from our workflow envisions that IgG molecules, being flexible in their solution environments, are coerced into semi-collapsed conformations by their shrinking droplets. This theory is supported by an MD study which saw gradual desolvation of ubiquitin, cytochrome c and holo-myoglobin >100 ns trajectories[42]. While IgG semi-collapsed conformations are accessible through Fab arm sampling, this procedure provides two additional benefits. Firstly, computational timescales can be cut significantly as the IgG has already collapsed to mimic a state in which nearly all of the solvent molecules have already evaporated from the protein. As a result, our gas phase MD simulations converge after a much shorter period of time. Secondly, charges cannot be mis-assigned to surfaces of the protein which later form the collapsed interfaces (which would prevent collapse). Therefore, we speculate that CRM charge transfer occurs concurrently or after partial collapse resulting in charge migration from solvent to exposed protein surfaces. This hypothetical model has been summarised in Figure 2.4.

In summary, probing the conformational dynamics of antibodies by IM-MS has led to several interesting conclusions due to their intrinsic flexibility. Firstly, the flexibility of IgG molecules can be represented through Fab arm sampling and allow inferences to be made about their solution dynamics. Building on these solution-relevant conformations, we theorized that IgG molecules undergo partial collapse in solution which eventually leads to their collapsed topologies in the mass spectrometer. The ability to model these collapsed structures accurately has provided insight into the experimental *CCS* distribution of these flexible molecules.

Overall this study highlights the need for a predictive model when interpreting gas phase data as non-globular proteins are unlikely to retain their native structures. A combination of high throughput IM-MS and molecular modelling may therefore provide complementary structural and dynamical information to other biologically relevant techniques such as hydrogen deuterium exchange mass spectrometry. Such approaches may facilitate invaluable data interpretation which might not be possible without structural representations. Ultimately, we anticipate that this workflow will be applicable to other flexible proteins currently eluding solution or gas phase structural and dynamical characterization.

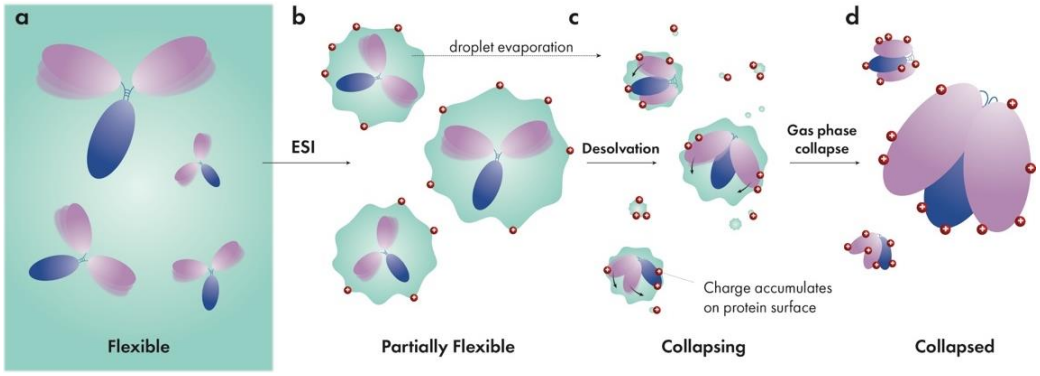


Figure 2.4. Proposed collapse pathway of IgG during ESI. (a) IgG molecules exhibit full flexibility in solution. (b) Nanospray ESI produces charged droplets in which IgG molecules retain partial flexibility depending on droplet size. (c) Gradual evaporation of droplets coerces flexible IgG molecules into more compact topologies. Solvent charges migrate to protein surfaces as they become exposed through desolvation (CRM). (d) Dry protein ions are inflexible in vacuum and represent a distribution of compact conformations.

Chapter 3: Deuteros 2.0: improved software for rapid analysis and visualisation of data from hydrogen deuterium exchange-mass spectrometry

Preface

The work presented in this chapter relates to the following article:

Lau, A. M. C., Ahdash, Z., Martens, C., Politis, A. (2019). Deuteros: software for rapid analysis and visualisation of data from differential hydrogen deuterium exchange-mass spectrometry. *Bioinformatics*, btz022, https://doi.org/10.1093/bioinformatics/btz022.

In early 2019, the first iteration of our Deuteros software was published. Since then, the code of Deuteros has been significantly reworked and upgraded in a second iteration - Deuteros 2.0. The 2.0 version of Deuteros constitutes a major change in both the code, interface and robustness of the data handling methods performed by the software. In this chapter, the workflow of Deuteros 2.0 will be described. The publication of the original Deuteros can be found in the Appendix. Data of the CSN-CRL2 complexes can be found in Chapter 4: of this thesis.

Author contributions

As first author of the above article, I conceptualised, designed and developed all code for *Deuteros* and *Deuteros 2.0*. I also led the writing of the manuscript, generated all figures, maintain the GitHub repository and coordinated the submission of the Deuteros article to *Bioinformatics*. Miss Zainab Ahdash and Dr Chloe Martens either aided in workflow design or contributed data for testing purposes. Dr Argyris Politis assisted with software suggestions.

3.1 Background

3.1.1 The HDX-MS Pipeline

HDX-MS data analysis begins where acquisition ends. Although data analysis can be performed manually, this is a time-consuming process and many pieces of software have been developed over the last several decades to streamline this process. Each software employs unique algorithms which intake data from HDX-MS and provide an output that can be interpreted for biological importance. While each software is different in design and their handle of the data, the typical stages of analysis are: 1) identification of peptides - generation of a peptide list, 2) extraction of isotopic envelopes belonging to each peptide and measurement of masses, 3) data visualisation (Figure 3.1).

Figure 3.1. From machine to biology: typical steps of HDX-MS.

The first stage involves the generation of a list of peptides for the protein of study. In some HDX-MS workflows, such as that for Waters instrumentation, this list can be generated through identification of experimentally observed peptides from LC-MS/MS, using proteomics platforms such as the Waters ProteinLynx Global Server (PLGS). For data collected with other setups or vendors such as the Thermo Scientific line of Orbitrap instruments, all-in-one data processing packages such as HDX Workbench[122] will perform the proteomics database search step within their automated pipeline. Other methods have also been developed which do not require

pre-established peptide lists such as AUTOHD and HEXICON[123,124]. These software instead use *in silico* digestion algorithms to generate a list of all possible peptides.

In the second stage, the clusters of isotopic envelopes belonging to each peptide identified in stage one is extracted and the corresponding mass determined. The mass of each deuterated peptide is also calculated. Similar to stage one, there are various methods of performing this step. Stage three represents a diverse range of different stylisations and representations that have been adopted for visualisation of HDX-MS data. These range from linear coverage maps that display the peptide coverage of the experiment, to two dimensional graphs that display the kinetics of each peptide. An optional fourth step is that statistical analysis can be applied to the peptide data to yield groups of peptides that behave in the same manner. A range of statistical approaches have been adopted for the treatment of HDX-MS peptide data, both in the context of single-state and differential comparisons. Conclusions can then be drawn about these peptide regions in the context of the biological system. In some software specialised for visualisation such as Deuteros, statistically treated peptide data can be visualised on molecular structures to further facilitate the interpretation of structural changes as a result of an environmental or conformational difference.

3.1.2 Software for HDX-MS data analysis

Since the advent of HDX-MS, the main bottleneck of the technique was in the speed that data could be collected for a set of biological samples. Since then, the commercialization of HDX-MS instrumentation and other efforts in enhancing automation for HDX, has shifted the rate-limiting step towards data analysis[122,123,125]. In short, data analysis for HDX-MS involves the extraction of peptide-sequence specific information from a set of raw MS files. This requires first identifying the sequence identity of a peptide, locating its relevant spectra and evaluating the peptide isotopic envelopes in order to determine the mass of the peptide. This is

repeated for all peptides at all timepoints, for all experimental conditions and replicates. Reviewing the peptide mass in the context of the deuterium labelling, provides a kinetic assessment of the peptide's behaviour in the folded protein. While these steps appear to be straightforward, in practice, data processing for HDX-MS is a laborious stage of the workflow, requiring significant manual effort from users to extract information from the raw data.

The typical peptide yield from an HDX-MS experiment is in the order of 10^2 peptides. Considering multiple labelling timepoints, measurement replicates and experimental conditions, the number of peptides quickly grows to unmanageable levels. For an example of how the number of peptides increases, consider the 14-subunit, ~500 kDa CSN-CRL2~N8 complex of which we recently characterized through HDX-MS. On average, 50 peptides per subunit may be expected from the complex. Taking into account four timepoints, technical replicates of $n = 3$ and in a differential comparison, results in 16,800 assignments across the dataset. It is unsurprising that without assistance from data processing software, HDX-MS of large or compositionally heterogeneous systems may be entirely impossible to tackle using this method. As a consequence, the last 20 years has seen the development of a wide variety of computational methods that can assist with data processing. These tools fulfil a myriad of niches, ranging from data processing of raw MS files, statistical analysis and data visualisation. These software can be broadly classified as doing one or more of the following: a) data extraction, b) statistical analysis, c) data visualisation. The release dates, publications and category of each software or related methodology is listed below in Table 3.1. The following sections will explore some concepts seen across HDX-MS software.

Table 3.1. Software for HDX-MS data analysis

Year of Release	Software Name	Developer & Publication	Category
2001	AUTOHD	Palmblad *et al*. 2001 [124]	Data extraction
	DXMS	Woods & Hamuro, 2001 [126] *with* Sierra Analytics (Modesto, US)	Data extraction
2006	HX-Express	Weis & Engen, 2006 [127]	Data extraction
2007	The Deuterator	Pascal *et al*. 2007 [128]	Data extraction
2008	TOF2H	Nikamanon *et al*. 2008 [129]	Data extraction
2009	HD Desktop	Pascal *et al*. 2009 [130]	All-in-one
	Hydra	Slysz *et al*. 2009 [125]	Data extraction
2010	MSTools	Kavan *et al*. 2010 [131]	Data visualisation
	HeXicon	Lou *et al*. 2010 [123]	Data extraction
2011	HDX-analyzer	Liu *et al*. 2011 [132]	Data extraction and statistics
	ExMS	Kan *et al*. 2011 [133]	Data extraction
	DynamX	Waters Corporation (Milford, US)	All-in-one
2012	HDXFinder	Miller *et al*. 2012 [134]	Data extraction
	HDX Workbench	Pascal *et al*. 2012 [122]	All-in-one
2014	MS Studio	Rey *et al*. 2014 [135]	All-in-one
	HeXicon 2	Lindner *et al*. 2014 [136]	Data extraction and visualisation
	HDExaminer	Sierra Analytics (Modesto, US)	All-in-one
2016	MEMHDX	Hourdel *et al*. 2016 [137]	Statistics and visualisation
2019	Deuteros	Lau *et al*. 2019 [138]	Statistics and visualisation
	HDX-Viewer	Bouyssie *et al*. 2019 [139]	Data visualisation

3.1.3 Methods for determining peptide mass

The extraction of HDX-MS data consists of the steps necessary to identify the isotopic envelope or distribution belonging to a particular peptide, and from this distribution, extract features that allow interpretations to be made. The isotopic envelope is the "information unit" of HDX-MS data and numerous useful parameters can be calculated, such as the mass of the peptide, amount of deuterium incorporation and the distribution width. There are two primary methods for determining the mass of a peptide from HDX-MS files. These will be explored in the following sections. The width of the isotopic envelope is a useful feature that can be used as a diagnosis of dynamics within the EX1 or EX2 exchange regimes. It is worth mentioning that most if not all of the software listed in Table 3.1 assume that residues within each peptide follow EX2 dynamics.

Determining peptide mass can be done through one of two types of methods: I) via determination of the distribution centroid m/z; or II) by fitting a theoretical isotopic envelope to the measured data. Using the centroid m/z method, the distribution centroid is calculated as the intensity-weighted average of the peaks belonging to the isotopic envelope. This, however, typically assumes EX2 exchange (unimodal) unless measures are set up to detect the distribution shape. In contrast, the isotopic envelope fitting methods first generate a theoretical distribution based on a known peptide composition identified earlier through the proteomics stage of the HDX-MS workflow. The theoretical distribution is then fitted and minimised to the measured distribution via linear regression. The fitting adjusts a component that accounts for the shift in mass and change in distribution shape, allowing the level of deuterium incorporation to be calculated. Since the shape of the isotopic distribution is an important factor for the goodness of fit, this method requires the distribution to be resolvable.

3.1.4 AUTOHD: mass determination through isotopic envelope fitting method

The necessity of computational assistance in data processing is visible in the timeline of publications detailing software developments for HDX-MS. In 2001, the HDX-MS community saw the release of the first data extraction software: AUTOHD and DXMS. DXMS, developed by Virgil Woods in collaboration with Sierra Analytics (Modesto, US), was offered as a commercial solution[126]. The early collaboration between Woods and Sierra Analytics, would later develop into the HDExaminer all-in-one platform. AUTOHD developed by Palmblad *et al*. was the first to apply the theoretical fitting method in an automated manner for HDX-MS[124]. AUTOHD is a command-line based software that employs a Fourier deconvolution method for generating theoretical isotopic envelopes. The deconvolution method was unique in that it recognised that the isotopic envelope of a peptide, can be generated through convolution of the individual isotopic envelopes of the constituent elements of the peptide. In other words, if the identity of a peptide is known, its isotopic envelope is a summary of all of its constituent elements. Calculating the isotopic envelope of peptide $a_{peptide}$, involves taking the elementwise product of the Fourier transformed (F) isotopic envelopes for each H, C, N, O and S elements, (denoted as a_X for each element and n_X is the number of element X) and then the inverse Fourier transform (F^{-1}) to recover a convoluted distribution for $a_{peptide}$ (3.1):

$$a_{peptide} = F^{-1}(F(a_H)^{n_H} \cdot F(a_C)^{n_C} \cdot F(a_N)^{n_N} \cdot F(a_O)^{n_O} \cdot F(a_S)^{n_S}) \qquad (3.1)$$

The most interesting feature of AUTOHD compared to its modern counterparts such as Waters DynamX, is that peptide mass determination using this method is performed independent of the undeuterated control[124]. Palmblad *et al*. recognised that in HDX, hydrogens can be divided into two categories: I) those that undergo

exchange too quickly or too slowly (i.e. hydrogens of side chains); or II) those that undergo measurable HDX (i.e. backbone hydrogens). Since group I hydrogens are either in equilibrium with deuterium in the surrounding labelling buffer, or do not exchange, their contribution to the calculated isotopic envelope can be accounted for through a $F(a_{solvent})^{nsolvent}$ term or, are already accounted for by $F(a_H)^{n_H}$. In contrast, the contribution of group II hydrogens - those with measurable deuterium uptake, are accounted for using the $F(a_{labile})$ term (3.2).

$$a_{peptide} = F^{-1}(F(a_H)^{n_H} \cdot F(a_C)^{n_C} \cdot F(a_N)^{n_N} \cdot F(a_O)^{n_O} \cdot F(a_S)^{n_S} \cdot F(a_{solvent})^{n_{solvent}} \cdot F(a_{labile}))) \quad (3.2)$$

Using this method, AUTOHD first generates a list of all possible peptide sequences given the experimental enzymatic setup. Using (3.2), a theoretical isotopic envelope is generated for each candidate peptide. Through a series of steps, isotopic clusters are identified and fitted with the theoretical envelope of each candidate peptide. The best fitting theoretical envelope is selected and minimised against the measured envelope to derive the level of deuterium incorporation. While this method is powerful in its ability to determine the level of deuterium incorporation without the need of a reference mass, it is worth noting that this method works well for protein sequences in the low molecular weight range (~15 kDa). The accuracy of AUTOHD quickly decreases as a function of molecular weight, presumably due to the dramatic increase in number of theoretical candidate peptides generated for longer sequences[124].

3.1.5 HX-Express: mass determination through the centroid *m/z* method

HX-Express was released in 2006 by David Weis and John Engen[127]. Rather than a standalone software package or command-line scripts, HX-Express is a collection of macros implemented in Microsoft Excel. HX-Express calculated peptide mass using the centroid m/z method and additionally automated the calculation of the

distribution width. Compared to AUTOHD which is executed via the command-line, the implementation of HX-Express in Microsoft Excel, while simple, was highly accessible to researchers due to both the common occurrence of Microsoft Excel and lack of specialised skills necessary to use the program. HX-Express provides three notable functions: 1) isotopic peak identification, 2) peak width determination and 3) centroid calculation. The algorithm of HX-Express takes as input, spectral data and other search parameters used to then search for isotopic clusters within an m/z range. The peak boundaries of a distribution are then determined and peak intensities I and m/z of each member peak i, are used to calculate the intensity-weighted average m/z using (3.3). Peak widths are calculated at specified percentage intensities of the distribution's height (e.g. 20% or 50%) and can be used for quantification of EX1 exchange[127].

$$m/z_{centroid} = \frac{\sum_i (m/z)_i \times I_i}{\sum_i I_i} \tag{3.3}$$

3.1.6 The Pascal series of HDX-MS software

In 2007, the HDX community would see the beginning of a line of programs that were eventually capable of performing data extraction, statistical analysis and data visualisation. These were 'The Deuterator', HD Desktop and HDX Workbench developed by Bruce Pascal *et al*. in the lab of Patrick R. Griffin. In 'The Deuterator', Pascal *et al*. emphasised the importance of providing easy access to data processing, both by implementing their program as a web-based application and its ability to accept input files in the common XML format from a range of high- and low-resolution MS instrumentation[128]. 'The Deuterator' uses both centroid m/z and theoretical isotopic envelope fitting methods for determining peptide mass allowing for robust high throughput data analysis. A significant drawback, however, was that no facilities post-extraction for data analysis, visualisation or any statistical analysis

were possible, meaning additional downstream software were still necessary for users to gain an understanding of their data. These areas of disadvantage were resolved when Pascal *et al*. released HD Desktop two years later in 2009, as the successor to 'The Deuterator' (albeit with a less catchy name)[130]. The theme of software accessibility was also inherited, implementing HD Desktop as an online web server and allowing data input from multiple instrumental sources. HD Desktop also represents a significant improvement on 'The Deuterator', with means of performing a myriad of data visualisation operations. Both single-state and multi-state experimental setups were possible, as were statistical analysis of the ensemble data. Finally, the output data could be represented on molecular structures through the use of Jmol[a].

HD Desktop was the answer of Pascal *et al*. to provide an all-in-one automated solution for data processing. The implementation of HD Desktop as a web server, increases accessibility to the software as it requires no local installation or any other software dependencies. Uploading data to a remote server, however, may be off-putting to some users who wish for their data to remain confidential. Pascal *et al*. addresses this issue with HDX Workbench, the third and latest iteration of their software line[122]. Released in 2012, HDX Workbench builds on the authors' five years of experience in handling and developing HDX-MS software. Pascal *et al*. rationalise that due to advances both in processing methods and computational power, there is no longer a need for data analysis to be performed using specialised servers[122]. As such, rather than being implemented as a web server, HDX Workbench is now installed either on a local computer or server, allowing greater file sharing capability between small groups of users. Another significant change is that raw files can be directly input into HDX Workbench as the *MSFileReader* library from Thermo Scientific is included. This, however, also reflects that at the time of release, only the

a http://www.jmol.org/

Thermo Scientific line of MS instrumentation is supported. The typical workflow for data analysis using HDX Workbench is shown in Figure 3.2.

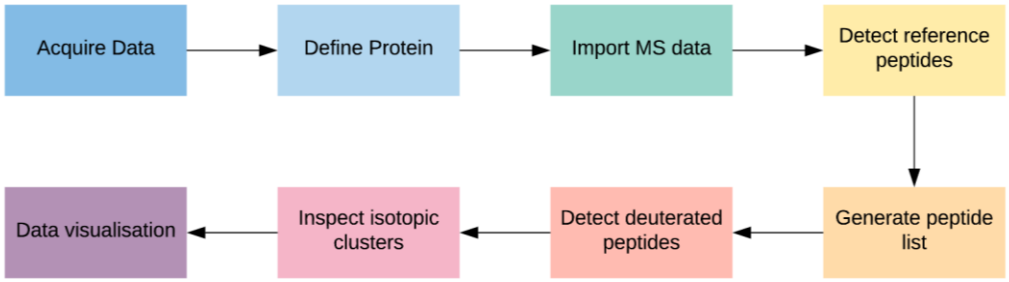


Figure 3.2. Data processing workflow of HDX Workbench. Adapted from Pascal *et al*. 2012 [122].

The hallmark of HDX Workbench is in its ability to analyse, organise and visualise large sets of HDX-MS data, directly from the raw input files. Data can be visualised in one of several styles, including coverage and heat maps which provide an overview of experimental coverage and deuteration levels of the entire dataset. The 'Experiment Comparison Tool' is an interface that displays HDX information as a function of experimental conditions and provides statistical assistance of datasets. Comparisons can be made between both sets of replicate data and between experimental conditions. For comparing replicate data, Pascal *et al*. employ analysis of variance (ANOVA) analysis to determine if the mean deuterium uptake over the replicate sets are significantly different. If a difference is found, a t-test is applied to quantify the significance through its P-value. Comparisons of experimental states instead employ a t-test for determining statistical significance. Results can then be visualised using a number of different tools including projection onto 3D model structures. The processed data is also archived for the user and can be reviewed at a later stage without the need for the software. Overall, the software line developed by Pascal *et al*. are a well thought out series of programs that addresses many bottlenecks and inconveniences of HDX-MS data analysis.

3.1.7 MS Studio

In 2014, the Schreimer group released MS Studio as another alternative data analysis package for HDX-MS[135]. Like the Pascal *et al*. software, MS Studio was a reimagined and redesigned all-in-one package stemming from another software that the Schreimer group released in 2009 - Hydra[125]. While Hydra was specialised designed to perform for statistical analysis specifically from HDX-MS, the MS Studio has expanded to perform data analysis for an umbrella of different MS labelling techniques. The inputs and outputs of HDX Workbench and MS Studio are similar - the user provides raw data files directly to the software, isotopic clusters are automatically identified, extracted and analysed, deuterium uptake determined, and data visualised. Masses are determined via the centroid *m/z* method; however, the studio also possesses an advanced interface that allows users to manually curate the isotopic envelope and select peaks that belong to the distribution. MS Studio employs t-tests and P-value filtering to identify peptides which show statistically significant differences in differential HDX-MS. The peptide ensemble and the statistical significance of each peptide is visualised via a specialised plot, named the 'Woods' plot as homage to the late Virgil Woods who was one of the first pioneers of automated HDX-MS. The Woods plot displays the residue number against ΔDeuterium uptake for each peptide, where the length of each peptide corresponds to the length of the bar. A statistical filter can then be applied to the data to identify significant peptides. Those with positive ΔDeuterium uptake, i.e. become 'deprotected' or 'destabilised' as a result of a change in the environment, are coloured red, while the opposite is coloured blue. This style of identifying potentially interesting peptides through applying statistical methods to peptide ensembles was also demonstrated by Houde, Berkowitz and Engen earlier in 2011[140]. Houde *et al*.

utilised a similar 'mirror' plot which was analogous to the Woods plot but used a numerically ordered peptide index in place of residue number (Figure 3.3). Applying a global statistical filter to the data, provided a convenient method of quickly identifying peptides of potential interest.

A major feature of the studio is its integration with HADDOCK which allows for the modelling of protein-protein and protein-ligand interfaces using the statistically relevant peptide data from HDX-MS[135]. This feature makes MS Studio unique in the all-in-one category of software since it is the only one that includes model building activities. MS Studio identifies a set of residues statistically determined via HDX-MS to be involved in interactions between proteins and other proteins or their ligands. These residues are passed to the HADDOCK server which performs the in silico docking. One HADDOCK run is performed for each residue identified and the results are clustered to identify possible conformations of the interaction.

3.1.8 DynamX

Packaged alongside its line of Synapt QTOF HDX-MS instrumentation, the Waters Corporation (Milford, US) released the DynamX software as an all-in-one to complement data analysis from HDX-MS. DynamX is the only software in the all-in-one category to be developed and released by an MS vendor, and is designed to be used together with another Waters developed software. A major limitation of DynamX is that it is vendor specific to only Waters instruments. The ProteinLynx Global Server (PLGS) is a proteomics platform used to identify peptides from LC-MS/MS data and unpacking this data for input into DynamX. In the first step of data analysis, 'ion-accounting' files containing a list of peptides identified through PLGS are input into DynamX along with the raw HDX-MS files for both undeuterated and deuterated reference files.

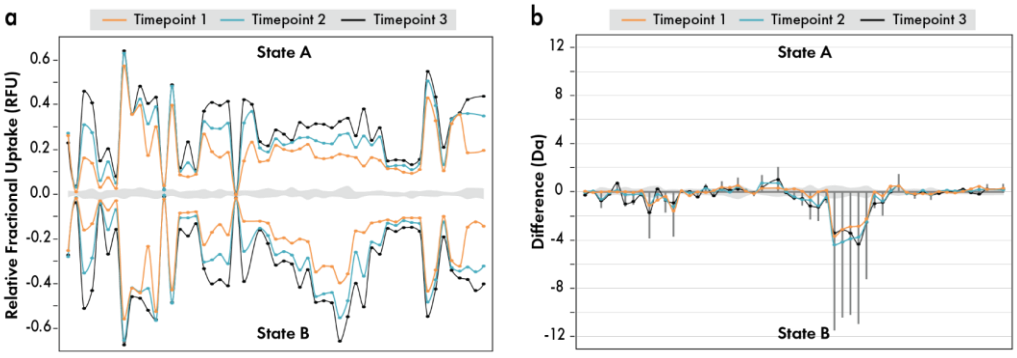


Figure 3.3. Representation styles for differential HDX-MS from DynamX software. (a) Butterfly plot from Waters DynamX. States A and B are shown on the top and bottom halves of the plot, mirrored along $\boldsymbol{RFU = 0}$. Each point on the plot is a single peptide. Orange, teal and black represent different timepoints. The grey trace along $\boldsymbol{RFU = 0}$ represents the 'error band' from DynamX. (b) Difference plot from Waters DynamX. Data shown is identical to (a). The grey bars represent the sum of differences over all timepoints. Both plots have been edited using Adobe Illustrator to increase legibility.

The peptide ensemble is then filtered for reproducibility and used to populate a list of peptides in the software's main interface. For each peptide, DynamX identifies an *m/z* range that the isotopic envelope can be found in. The software offers a streamlined interface for the manual curation of peak assignments, allowing users to check and amend incorrect assignments if necessary. There are no statistics available. Post assignment, the data can be visualised in several methods, including kinetic plots for individual peptides, coverage and heat maps and 'butterfly' and 'difference' plots (Figure 3.3). The software also provides the option to output peptide uptake data to molecular structures using PyMOL. The name 'butterfly' derives from the appearance of the plot being similar to the symmetric wing profile of a butterfly when two plots are simultaneously shown as in the multi-state format. The difference plot is a condensed plot format that calculates the difference between peptides shown in the butterfly plot. A useful feature of these plots is their inclusion of replicate standard deviation within the plots in the form of a grey silhouette along y=0.

3.1.9 MEMHDX

MEMHDX developed by Hourdel *et al*. in 2016, is a standalone software for analysis of differential HDX-MS[137]. While MEMHDX is not an all-in-one software, its features are heavily geared towards statistical analysis. Unique to MEMHDX, is that Hourdel *et al*., acknowledge that the currently seen statistical methods in HDX-MS data are only carried out at certain acquisition timepoints rather than holistically. MEMHDX uses a 'Mixed-Effects Model' in which variation over replicate datasets are treated as random effects, while also account for time dependency of the measurements. The focus of MEMHDX is on differential comparison of experimental states. For each peptide comparison, MEMHDX calculates two P-values: P-value of the 'magnitude of ΔDeuterium uptake' and P-value of the change in dynamics. The P-values of each peptide comparison used to cluster peptides into groups showing similar uptake kinetics. Results are displayed in a 'Logit plot' which plots the P-value of ΔDeuterium uptake against the P-value of change in dynamics (Figure 3.4). Interestingly, MEMHDX determines peptide dynamics to be in one of four categories for each differential comparison: those that experience upon ligand binding 1) increased dynamics, 2) decreased dynamics, 3) no change in dynamics and 4) increased dynamics in both states. These groups of peptides are coloured in red, blue, grey and green in the Logit plot respectively.

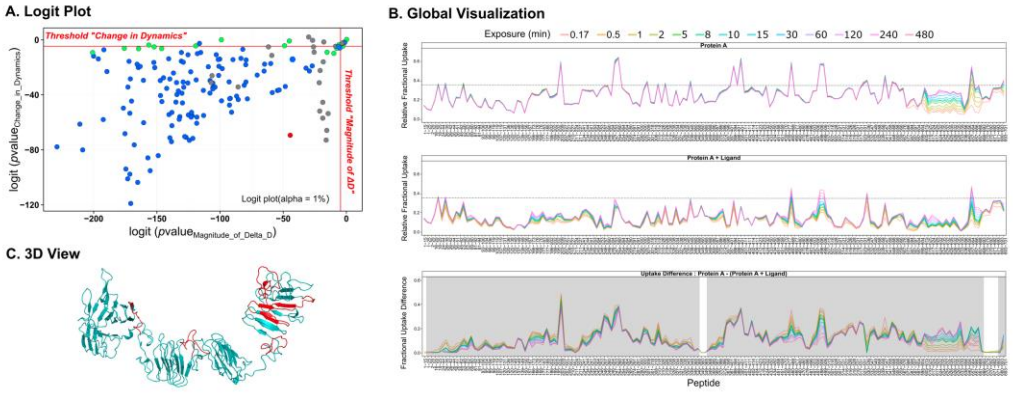


Figure 3.4. Example of outputs styles from differential HDX-MS using MEMHDX. Data shown are for the comparison of the apo CyaA toxin and when bound to a ligand. (A) Logit plot displaying P-value of change in dynamics against P-value of magnitude of ΔDeuterium uptake is shown for peptide comparison. Colours indicate clusters of peptides that exhibit similar behaviour. Red/blue peptides show increased dynamics in the holo or apo states respectively. Grey/green are peptides which show dynamics in either both states and no dynamics respectively. (B) Global visualisation plot allows users to visualise the kinetics of deuterium uptake. Top and middle panel show the relative fractional uptake of peptides in apo and holo states. Bottom panel shows a differential comparison of the two. (C) Structural representation of results. Cyan and red colouration indicate regions of dynamics and no dynamics upon binding of the ligand. Taken from MEMHDX website at http://memhdx.c3bi.pasteur.fr/).

3.1.10 Next steps in HDX-MS data analysis

As seen in the above sections, many software have been developed for the analysis HDX-MS data. These software range from those that perform single steps in the data analysis workflow, such as extraction of isotopic envelopes (e.g. AUTOHD), or statistics (e.g. MEMHDX), while others aim for true automation of the workflow through all-in-one packages (e.g. HDX Workbench, MS Studio, DynamX). The flexibility of recent all-in-one software to support vendor or instrumentation-independent data processing, has also vastly improved the accessibility of HDX-MS to researchers. These advances in software developments have led some in the community to believe that data analysis is no longer the bottleneck of the technique[122]. Downstream to data analysis, however, is the issue of data interpretation and whether or not this is meaningful. Although data can be acquired

and processed in a semi-high throughput manner, the appropriate interpretation of HDX-MS data is still paramount to gaining correct biological insights into the system of study. The increase in acquisition and processing efficiency has also been accompanied by an increase in the complexity of systems being studied, such as membrane protein systems[141,142], intrinsically disordered proteins[143] and macromolecular complexes[135,144,145]. The current statistical methods employed in HDX-MS are also used only as data filters for qualitatively clustering peptides into categories (e.g. deprotected, protected or no change), meaning that both the rate of change and magnitude of deuterium uptake difference are effectively ignored once the filtering threshold is crossed. In other words, current methods lack the ability to fully extract the information that HDX-MS offers.

In summary, although current software developed over the last few decades have been indispensable for streamlining HDX-MS data analysis, future software must be able to provide accurate and meaningful methods of interpreting HDX results in the context of protein dynamics. This should be done in relation to both single-state and differential modes of HDX-MS. In the following sections of this thesis, we will showcase the development of our HDX-MS visualisation software Deuteros, as well as the significant increases in software features following the first publication of the software.

3.2 Aims & Objectives

In this chapter, we aim to develop a data analysis and visualisation software that is capable of supporting interpretation of HDX-MS data. While this role is already served by many existing software, we find that there are incompatibilities between these, such as HDX Workbench, and Waters-based HDX-MS instrumentation (Synapt G2-Si with DynamX). Moreover, our in-house HDX-MS setup relies on the use of PLGS for peptide identification and DynamX for mass assignment, effectively dismissing the use of other all-in-one solutions such as MS Studio. A combination of these circumstances means that the simplest way of supplementing our current set up without drastically redesigning our HDX-MS workflow, is the addition of statistical analysis and visualisation steps downstream of mass assignment.

Taking into account these several points, the data analysis software Deuteros was developed specifically to address the following aims:

1. To provide an easy to use platform for statistical analysis and visualisation of HDX-MS data
2. To provide a method of extracting meaningful information from HDX-MS data in the context of structural interpretations
3. To develop a foundation software that can be dynamically adjusted to the needs of the evolving HDX-MS field
4. To keep user interaction to a minimum where unnecessary
5. Free to use, offline and accessible on most computers

To meet these aims, the following objectives must be completed:

Objective 1: Develop a function for reading the output of DynamX with minimal manual edited needed

There are several data output formats offered by DynamX - these will be outlined in greater detail in section 3.3. As mentioned before, software such as MEMHDX require a degree of manual editing of the DynamX output file prior to data analysis. To improve accessibility to data analysis, we will avoid the need for any manual intervention through two actions: 1) the output of DynamX will be directly parsed into Deuteros. Any data that is not needed, will be simply ignored, rather than require they be manually removed. 2) Any required manual labelling, such as the number of replicates, will be performed programmatically in Deuteros, rather than by the user. Meeting this objective will significantly improve the speed and accessibility to data visualisation features offered by the software.

Objective 2: Develop methods of data visualisation, including 2D and 3D representations

There are three categories of visualisation styles:

1. Simple 2D maps, e.g. coverage and redundancy maps, etc.
2. Complex 2D maps, e.g. deuterium uptake plots, Butterfly plot, Difference plot, Woods plot, Logit plot, etc.
3. Visualisation on 3D structures, e.g. projection of deuterium uptake on PDB structures.

We will begin by implementing simple 2D maps that show peptide coverage and redundancy as these are typically the first types of data visualised. Other complex 2D maps such as the Woods plot should be added to provide facilities for performing differential HDX-MS analysis. Woods plots can also take advantage of established statistical methods for determining the significance of differential uptake[140]. In turn, peptides identified as showing statistically significant deuterium uptake differences can be mapped onto protein models using PyMOL.

Objective 3: Include the latest recommendations from the HDX-MS community

Deuteros should include recommended practices detailed in the recent HDX community guideline paper[146]. This involves including automated methods for back exchange correction, as well increasing transparency on the quality of HDX-MS datasets.

Objective 4: Design an accessible and simple to use graphical user interface

To meet this objective, we have opted to develop Deuteros using the application development suites of MATLAB. There are several reasons for using MATLAB. Firstly, the author's programming experience is primarily in the MATLAB language, meaning that a simple application can be crafted quickly. MATLAB provides facilities to streamline development of the graphical user interfaces (GUI) in two suites: GUIDE and Appdesigner. These will be covered in 3.3 Materials and Methods. Finally, although MATLAB is not free to use, applications packaged using MATLAB are independent and do not require that MATLAB be installed on the system. A requirement however is that the MATLAB Runtime library must be installed. The library is free to download and install from MathWorks[b].

Applications designed using the MATLAB Appdesigner are visually uncluttered, clear and appealing, making it suitable for designing Deuteros. The following sections of this thesis chapter will outline the development of Deuteros in the context of meeting these aims and objectives.

b https://uk.mathworks.com/products/compiler/matlab-runtime.html

3.3 Materials and Methods

3.3.1 Datasets

All experimental details of the CSN-CRL2 system can be found in Chapter 4 of this thesis. HDX-MS data for CSN and CRL2 complexes were collected by Dr Chloé Martens.

3.3.2 Code development in MATLAB

Deuteros is a standalone MATLAB graphical user interface (GUI) that is available to both Mac and Windows operating systems, providing it is capable of installing and running the MATLAB runtime library [c] (free to download from MathWorks Inc., Massachusetts, USA). The first *Deuteros* program was conceptualised as a standalone script that processed data specifically from differential HDX-MS which was pre-processed by DynamX software (Waters). Early prototypes of *Deuteros* featured no GUI but principally required two inputs - the so called '*state*' and '*difference*' files which could be exported from DynamX. Along with these files, the user would also have to supply the start and end residue numbers to the software. In addition to '*state*' and '*difference*' files, DynamX can further export peptide data in its '*cluster*' format. Given the importance of these file structures, it is worth briefly introducing each of the export types and their differences.

3.3.3 The '*Cluster*' format

The '*cluster*' file is the lowest level of data exported from DynamX. It contains a per-protein, per-state, per-timepoint, per-replicate and per-charge list of peptides. It does not contain deuterium uptake values. The major advantage of the '*cluster*' format is that it retains all datapoints for each replicate observation, allowing more

c https://www.mathworks.com/products/compiler/matlab-runtime.html

complex statistical analysis to be performed. There are 15 columns in a '*cluster*' file (Table 3.2).

Table 3.2. Format of the DynamX '*cluster*' file

Protein	Start	End	Sequence	Modification	Fragment	MaxUptake
CSN1_HUMAN	17	30	MQIDVDPQEDPQNA			10
CSN1_HUMAN	17	30	MQIDVDPQEDPQNA			10
CSN1_HUMAN	17	30	MQIDVDPQEDPQNA			10
CSN1_HUMAN	17	30	MQIDVDPQEDPQNA			10
CSN1_HUMAN	17	30	MQIDVDPQEDPQNA			10
CSN1_HUMAN	17	30	MQIDVDPQEDPQNA			10

MHP	State	Exposure	File	z	RT	Intensity	Center
2187.95	CSN	0.00	Ref1	2	5.32	1398163	1095.04
2187.95	CSN	0.00	Ref2	2	5.33	1394571	1094.99
2187.95	CSN	0.00	Ref3	2	5.32	1700971	1095.06
2187.95	CSN	0.25	15_1	2	5.26	792029	1099.59
2187.95	CSN	0.25	15_2	2	5.36	23151	1099.76
2187.95	CSN	0.25	15_3	2	5.26	620447	1099.54

Protein refers to the protein identifier used initially in PLGS for peptide search. *Start* and *End* are the start and end residue numbers of peptide *Sequence*. *Modifications* and *Fragments* are shown if found in the PLGS search. *MaxUptake* is the theoretical maximum deuterium uptake of the peptide, calculated by equation (3.4) where N_{res} and N_{pro} are the total number of residues and prolines in the peptide respectively.

$$MaxUptake = N_{res} - N_{pro} - 1 \tag{3.4}$$

MHP is the mass of the singly charged ion measured in Daltons. *State* and *Exposure* refer to the state name and exposure time designated by the user in DynamX during file import. *Filename* details the file that the replicate data has been sourced from. Typically, this is performed as a technical replicate. *z* is the charge of each peptide — more than one set of charges may be present in the *cluster* file. *RT* is the retention time of the peptide. *Intensity* refers to the signal intensity of the peptide. *Center* represents the centroid *m/z* for the isotopic envelope of the peptide.

3.3.4 The '*State*' format

The '*state*' format is the next level upwards from the '*cluster*' format and contains largely the same headings. However replicate data is no longer accessible and have been aggregated into a single value along with standard deviations for each set of observations. An advantage of the '*state*' format over '*cluster*' is that deuterium uptake values are readily available to users. The method of calculating the average uptakes and standard deviations over the replicate data is also poorly described in the DynamX documentation and thus may be prone to errors from manual calculation. There are 16 columns in a '*state*' file (Table 3.3).

Table 3.3. Format of the DynamX '*state*' file

Protein	Start	End	Sequence	Modification	Fragment	MaxUptake	MHP
CSN5	15	21	LANNMQE			6	819.3665
CSN5	15	21	LANNMQE			6	819.3665
CSN5	15	21	LANNMQE			6	819.3665
CSN5	15	21	LANNMQE			6	819.3665
CSN5	15	21	LANNMQE			6	819.3665
CSN5	15	21	LANNMQE			6	819.3665
CSN5	15	21	LANNMQE			6	819.3665
CSN5	15	21	LANNMQE			6	819.3665

State	Exposure	Center	SD	Uptake	SD	RT
A	0.00	819.794	0.029	0.000	0.000	3.447
A	0.25	823.673	0.040	3.879	0.049	3.406
A	5.00	823.485	0.074	3.691	0.080	3.455
A	30.00	823.468	0.046	3.674	0.054	3.441
B	0.00	820.173	0.101	0.000	0.000	3.394
B	0.25	823.039	0.057	2.866	0.116	3.445
B	5.00	823.055	0.128	2.882	0.163	3.480
B	30.00	823.306	0.070	3.133	0.123	3.386

Protein, *Start*, *End*, *Sequence*, *Modifications*, *Fragments*, *MaxUptake*, *MHP*, *State* and *Exposure* are identical to the '*cluster*' file. In the '*state*' file, *Center*, *Uptake* and *RT* each represent the intensity-weighted mean over all replicate datapoints. The deuterium uptake (DU) of peptide i at time t (where $t \neq 0$) is calculated by (3.5) and (3.6):

$$mass = (center \times z) - MHP \tag{3.5}$$

$$DU_i^t = mass_i^0 - mass_i^t \tag{3.6}$$

To calculate the intensity-weighted mean *Uptake*, $\overline{x}^*$ for each peptide at each timepoint, two matrices are generated: x, the DU matrix, and w, the intensity matrix. The DU matrix is calculated by subtracting the DU_0 vector, n, from each DU_t vector, m, generating an $\boldsymbol{n \times m}$ matrix (3.7). The intensity matrix is generated as the cross product of int_0 and int_t for the same t as x.

$$x = [DU_{t,rep1} \quad \overset{m}{DU_{t,rep2}} \quad DU_{t,rep3}] - \overset{n}{\begin{bmatrix} DU_{0,rep1} \\ DU_{0,rep1} \\ DU_{0,rep1} \\ DU_{0,rep1} \\ DU_{0rep1} \end{bmatrix}} \tag{3.7}$$

$\overline{x}^*$ is then calculated as the sum product of x and w, over the sum of w according to (3.8). *Center* and *RT* are calculated in the same manner.

$$\overline{x}^* = \frac{\sum_{i=1}^{N} w_i \, x_i}{\sum_{i=1}^{N} w_i} \tag{3.8}$$

The intensity-weighted standard deviation for each replicate set is given in *Center SD*, *Uptake SD* and *RT SD* is calculated according to (3.9).

$$SD = \sqrt{\frac{\sum_{i=1}^{N} w_i \, (x_i - \overline{x}^*)^2}{\frac{(M-1)}{M} \sum_{i=1}^{N} w_i}} \tag{3.9}$$

Where N is the length of x and M is the length of w. The method of calculating the intensity-weighted standard deviation displayed here is not explicitly given in the documentation of DynamX but have been cross-checked with values exported from the '*state*' file.

3.3.5 The '*Difference*' format

Finally, the 'difference' format is one of the highest-level tabulated formats from DynamX. It can only be generated from a DynamX session which has been loaded with differential data and is accessed through interactively by copying the data to the clipboard via a right-click context menu spawned in the '*Butterfly Plot*' window of DynamX. The plot must also first be switched to show the '*difference plot*' representation in the plot settings menu, before the '*difference*' data can be copy-and-pasted into a spreadsheet editor. The difference data contains at least 5 columns, and one additional column for every non-zero/reference exposure time (Table 3.4).

Table 3.4. Format of the DynamX '*difference*' file

Sequence	Start	End	Modification	t_1	t_2	t_3
LANNMQE	15	21		-1.0132	-0.8093	-0.5412
IYKYDKKQQQEIL	28	40		-0.0678	-0.5945	-0.4155
AAKPWTKDHHYFKY	41	54		-0.1143	-0.0489	-0.0535
PWTKDHHYFKY	44	54		-0.1119	0.0044	0.1152
VMHARSGGNLEVMGLMLGKV	66	85		0.1417	-0.1459	0.0763

The '*difference*' format is a highly simplified version of the '*state*' file and only contains the *Sequence*, *Start*, *End*, *Modification* and ΔDU for each exposure time, between two user-selected states (expressed in Da). While the user can control which two states are used to generate the '*difference*' data, caution should be taken to select the two states in the correct order such that the polarity of the difference values lead to the intended interpretation. Typically, ΔHDX-MS performs comparisons as Δ(variable condition - control condition), and positive changes in deuterium uptake are interpreted as deprotective or destabilising changes, and negative deuterium uptake differences as protective or stabilising. Additionally, since neither the filenames, states, or the comparison direction is recorded in the difference file, the '*difference*' format is prone to mix-up, which complicates its usage.

3.3.6 Changes to the design of the original Deuteros

There are a few notable differences between the original Deuteros[138] and Deuteros 2.0. These will be briefly highlighted here.

3.3.7 GUIDE vs appdesigner

The most significant difference between the old Deuteros and Deuteros 2.0 is something that is not visible to the user. The original release of Deuteros was designed using the MATLAB *GUIDE* interface which while provided an easy to use and accessible method of coding the software for beginners, was limited in its programmatic features (Figure 3.5). In *Deuteros 2.0*, the software has been re-written using the MATLAB *appdesigner* interface which is superior in terms of code editing functionality and GUI design (Figure 3.6). The *appdesigner* is a separate application to the MATLAB main window and is a development environment that facilitates the production of professional applications by providing developers with easy to use tools for designing and planning their software. In practice, the *appdesigner* is much more efficient to write and troubleshoot code due to its all-in-one environment that allows both GUI design and coding, while *GUIDE* only performs the design element of the software.

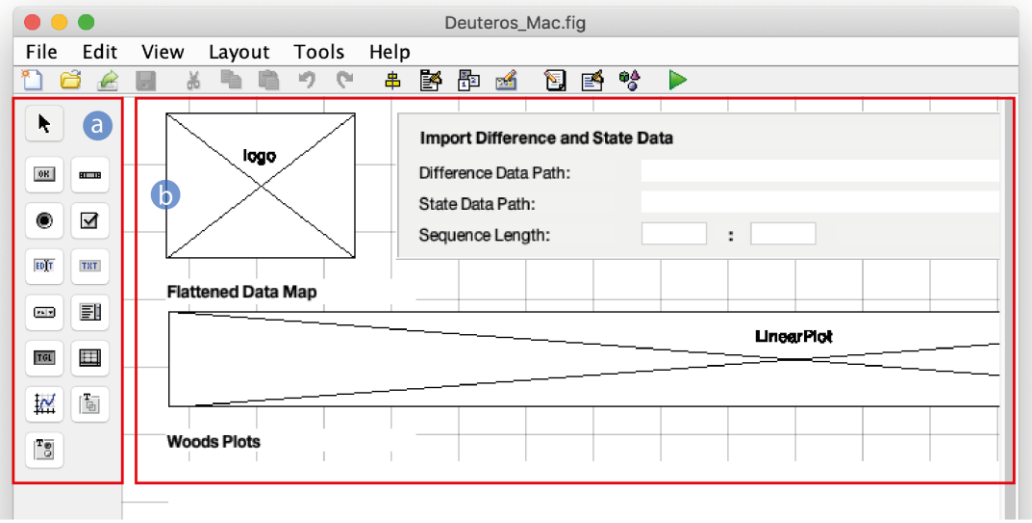


Figure 3.5. *GUIDE* app development interface for Deuteros. (a) The component palette of *GUIDE* allows UI elements (e.g. radio buttons, dropdown menus, axes, etc.) to be 'drag-and-dropped' onto (b), the layout editor for designing the Deuteros interface. *GUIDE* is only for designing the GUI and not directly for writing the underlying code.

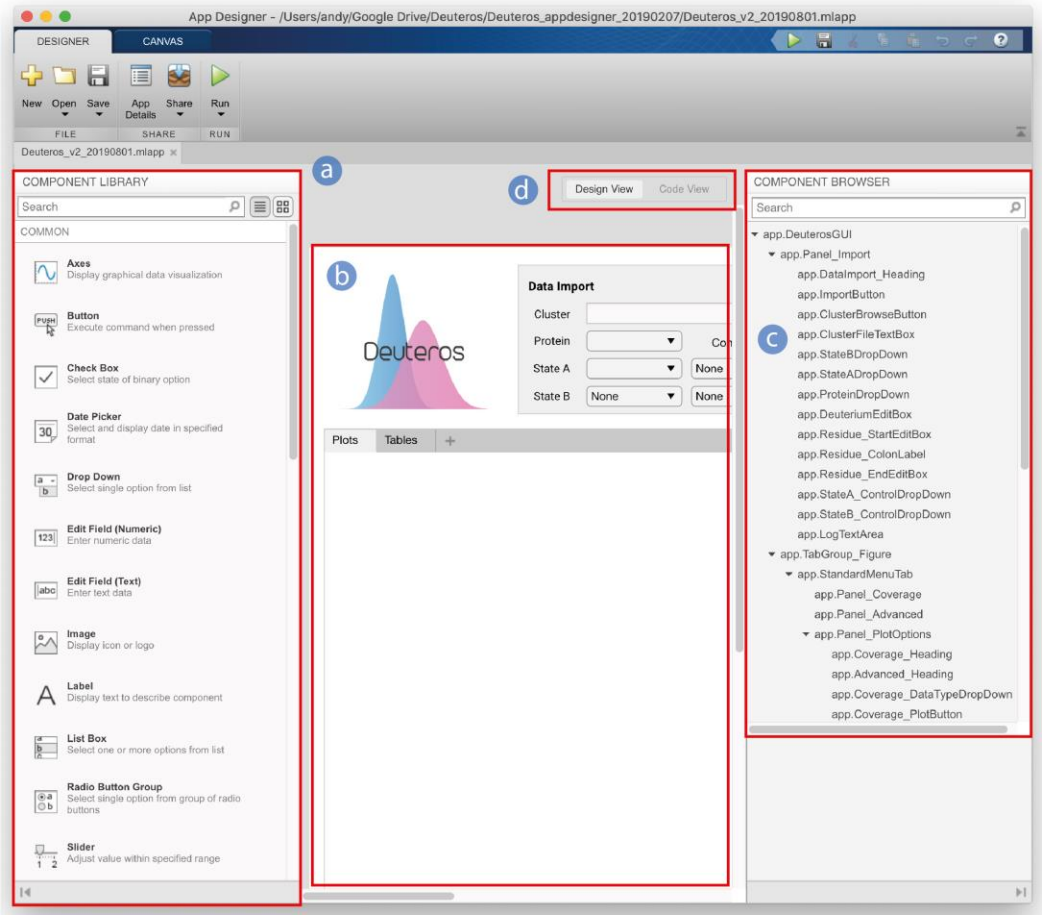


Figure 3.6. *appdesigner* development interface for *Deuteros 2.0.* (a) The component library of *appdesigner* operates in a similar manner to *GUIDE* but offers additional elements for developers. Components can also be 'drag-and-dropped' onto (b) the layout editor. The component browser (c) lists all UI components and their handles (unique name of the component used for accessing its functions, e.g. *app.ImportButton* refers to the import button on in the data import panel) and the panels that they belong to. Unlike *GUIDE*, *appdesigner* keeps track of the component list and any re-named handles, removing the need to constantly refer back to the code and manually update. (d) *Appdesigner* can be switched between design view and code views using the view toggle.

3.3.8 Cluster input

Firstly, a decision was made to migrate the import function of the new *Deuteros 2.0*, from using '*state*' and '*difference*' files to using instead the 'cluster' file format. The '*cluster*' file is superior in terms of the information content that it includes, providing

a breakdown of data on all proteins, states, timepoints and replicates. Meanwhile, the in the old Deuteros, differential uptake values were accessed directly from the 'difference' file which is both laborious to produce when a dataset contains more than one protein and state, contains a highly simplified version of a $\Delta(State_B - State_A)$ comparison, and lacks any measure of statistical variation across replicates used to compose the data. As a result, the '*state*' data - a less summarised version of the 'difference' file, but more summarised than the '*cluster*' format, must be provided by the user to compensate for the lack of replicate information. In reality, the '*state*' file alone can be used to calculate both the differential uptake (essentially rendering the 'difference' file useless) and provide statistical information. By requiring the user to supply both 'state' and 'difference' files was also cumbersome and prone to errors due to the large number of files that a user needed to keep organised in a typical dataset. Changing from ‘state’ and ‘difference’ files to ‘cluster’ input completes *Objective 1: Develop a function for reading the output of DynamX with minimal manual edited needed.*

3.4 Results

3.4.1 Deuteros 2.0 overview

As previously described, a typical HDX-MS experiment, from protein to interpretation, consists of four steps. These are 1) sample preparation and data acquisition, 2) peptide database search, 3) peptide mass assignment and 4) analysis and visualisation. Although not strictly enforced, Deuteros has been designed for data exported from Waters DynamX software. Deuteros adds a step 5 to the above workflow in that it performs downstream statistical filtering of peptides and enhances the repertoire of visualisation methods that is accessible to the user. Deuteros was designed with usability in mind and performs many layers of data processing behind the scene, to avoid unnecessary complications for the user. As a result, the graphical user interface (GUI) is clean and minimalist, and care has been taken to minimise the number of mouse clicks that a user needs to transform their HDX-MS data into graphical representations. This completes *Objective 4: Design an accessible and simple to use graphical user interface* described in 3.2 Aims & Objectives.

To improve accessibility, the application has been packaged into standalone software both for Mac and Windows operating systems using the MATLAB Compiler[d]. The MATLAB Compiler generates a Windows Installer executable (.exe) or Mac disk image (.dmg) which can be used to install the software as a standalone program. As mentioned previously, Deuteros and other MATLAB-coded applications requires the MATLAB Runtime Library as a prerequisite. The MATLAB Runtime Library can be downloaded free of charge from MathWorks Inc. and is used to run MATLAB applications without the need for MATLAB to be installed and a license active.

d https://www.mathworks.com/products/compiler.html

Installation of Deuteros is straight forward as there are no other software dependencies or options to select during setup.

3.4.2 Importing data to *Deuteros 2.0*

The GUI of Deuteros can be subdivided into several regions (Figure 3.7), in particular, the *Data Import* panel which handles the parsing of the HDX-MS data in the form of the '*cluster*' file.

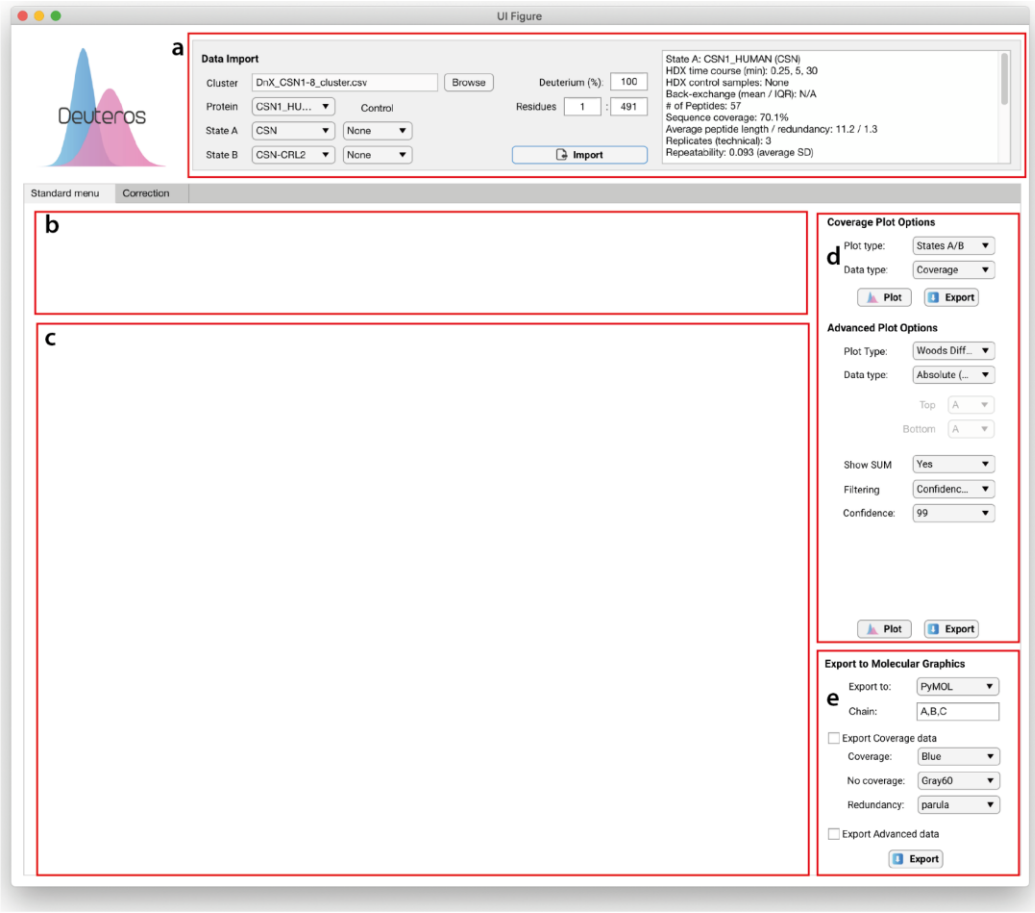


Figure 3.7. Deuteros 2.0 GUI. The GUI can be subdivided into five main regions: (a) The *Data Import* UI panel, (b-c) *Coverage Plot* and *Advanced Plot* UI panels, (d) *Plot Options* panel and (e) *Export to Molecular Graphics* options.

When a file is provided to Deuteros 2.0 via the *Browse* button, the program will generate a list of proteins and per-protein states which are used to populate the *Protein*, *State A*, *State B* and *Control* dropdown menus (Figure 3.8). There are no limits to the number of proteins and states that can be assigned to the dropdown menus. By default, the values of State B and Control dropdown menus are set to *None*, i.e. single-state analysis is performed. If multiple states are available, a second state can be selected in the *State B* dropdown to be used for differential analysis. The *Control* dropdown menus can be optionally set to a state housing back-exchange and in-exchange data which are used for data correction.

The *Browse* button has also been programmed to store in memory the path of the last file it accessed and to open at this path. To the right of the dropdown menus are edit boxes for Deuterium % and the *start* and *end* residue numbers. Deuterium % is the percentage deuterium content of the labelling buffer used during data acquisition. Deuterium % has been set to 100% by default. The value of is Deuterium % used to linearly scale the relative fractional uptake of the imported data. The *start* and *end* residue numbers are needed from the user to determine the length of the protein for both experimental coverage calculations and setting axis limits of plots. Although the '*cluster*' file (which contains the highest level of detail) is used, the user must manually provide the length of the protein as this information is not available in any of the outputs of DynamX. In reality, there are benefits in designing the residue numbers to be input manually. Firstly, this allows a simple level of customisation if the user is dealing with a protein construct which does not canonically begin from residue 1. Secondly, the user will be made more aware of the length of their protein. In our experience, this can lead to the discovery that the user's system has been truncated or the incorrect isoform of the sequence has been provided initially to PLGS. Finally, this method of inputting the residue numbers is less troublesome than requiring the user to supply a FASTA file of their protein of interest, especially as this would need to be changed every time the user selects a different protein. Requiring

a FASTA file would also not eliminate the need for the user to manually edit their file in the case of truncated proteins.

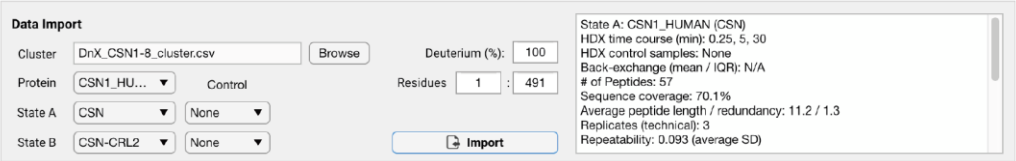


Figure 3.8. Data Import UI panel of *Deuteros 2.0*. On the left half of the import panel, interactive controls of cluster filename, protein, state A, state B dropdown menus, deuterium percentage and residue start, and end are found. Control states can also be selected using dropdown menus if back-exchange data is available for the selected state. On the right, a summary of the imported HDX-MS data for states A and B can be found. Metrics of this summary are those recommended by the HDX-MS community.

After the *Data Import* form has been completed, the *Import* button is used to begin importing the '*cluster*' data into the software. The *Import* button performs a significant number of important processing steps. In the first step, the protein and states selected from the dropdown menus by the user are used to filter the '*cluster*' file so that only the relevant peptides are processed. Next, four different MATLAB data tables are generated:

1. **StateA.table** - all State A peptides
2. **StateB.table** - all State B peptides
3. **Common.table** - Common peptides between States A and B
4. **Difference.table** - Difference between common peptides of State B - common peptides of State A

The four data tables are necessary due to the different representation styles that Deuteros 2.0 generates downstream of data import. *StateA.table* and *StateB.table* each houses a filtered list of peptides from each state. Unique peptides found in each state are preserved. *Common.table* is a concatenation of *StateA.table* and *StateB.table*, however only peptides that are common between States A and B are preserved. The *intersect* function (returns data common between two states) is additionally employed at the *exposure* time level to ensure that all exposures for each peptide match between the two states. Finally, *Difference.table* is generated

from iterating through a list of peptides and exposures in *Common.table*, and calculating the difference in *Uptake* between States A and B. The comparison is always performed as $\Delta(StateB - StateA)$.

Each of the four data tables are passed through an *aggregate* function which is used to convert the '*cluster*' style data to a '*state*' style table. The *aggregate* function calculates the intensity-weighted mean *Uptake*, *RT*, *Center* and corresponding standard deviation across all replicates for each peptide at each exposure time. In addition, the *RFU* of peptide i at time t is calculated from the *Uptake* using equation (3.10).

$$RFU_i^t = \left(\frac{Uptake_i^t}{MaxUptake_i^t}\right) Deuterium\% \tag{3.10}$$

As the '*cluster*' file retains all charge states that have been assigned to each peptide in DynamX, the *aggregate* function has been designed to perform an important sanitisation step which filters out all but one set of charge states for each peptide. This is necessary because mass assignment in DynamX requires users to manually inspect and assign the spectral peaks for every protein, state, peptide and timepoint within their dataset. It is part of the mass assignment procedure that only one set of charge states are assigned for each peptide, meaning that users must also manually unassign all other charge states. In practice, many charge states are often accidentally left assigned and thus will affect the calculation of deuterium uptake for those peptides. This highlights another advantage of accessing the HDX-MS data via the '*cluster*' format as it preserves all charge states that are even partially assigned, allowing these to be filtered out to provide more accurate $Uptake$ and $\Delta Uptake$ measurements. If multiple sets of charge states are found, the *aggregate* function will automatically remove all redundant sets based on their *Uptake SD* (Figure 3.9). Table 3.5 shows the format of cluster data following the aggregation function. In the modified cluster table, N the number of replicates that the $Uptake$ is calculated from, and $Uptake\_Reps$ the replicate data are preserved. These values are needed in

later stages for statistical analysis of the ensemble and also to allow the user to access the underlying data more easily if needed since this is not available by default from the *Cluster* file.

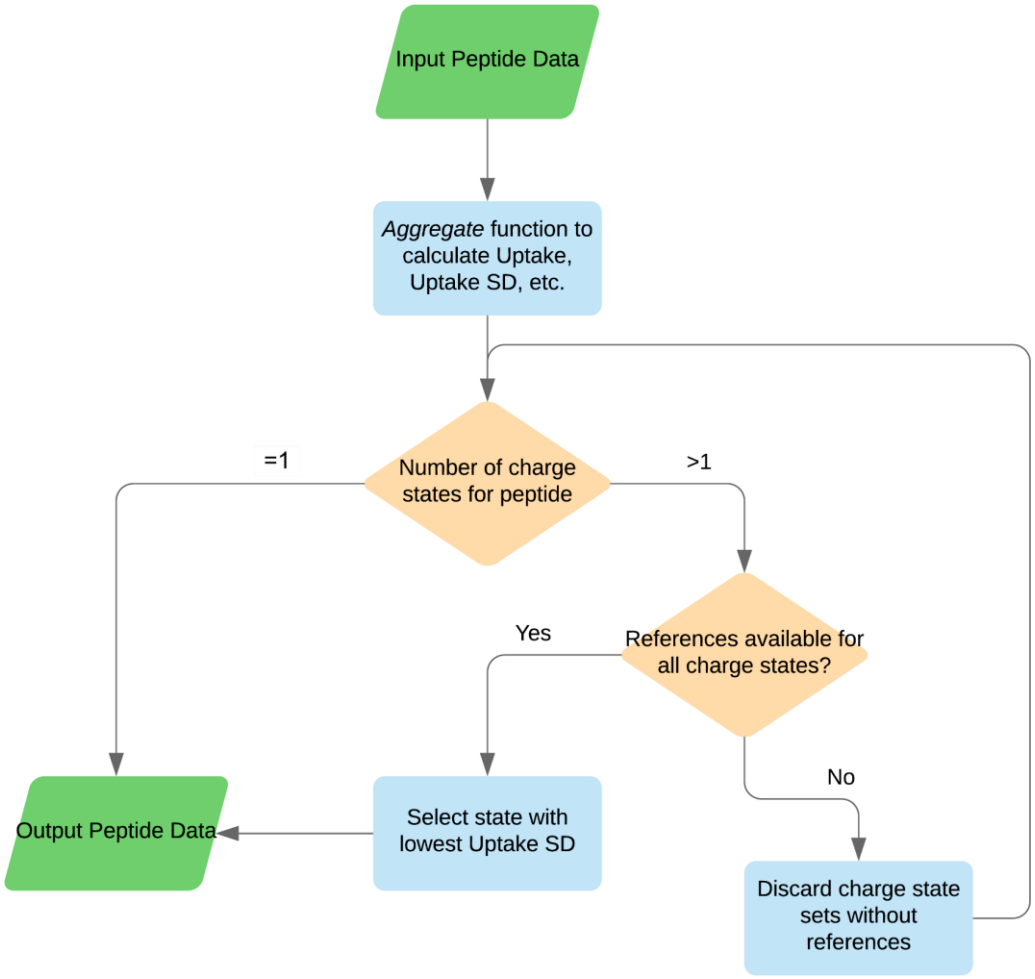


Figure 3.9. Flowchart for peptide charge state removal. Decisions, processes and outputs are represented by diamonds, rectangles and parallelograms respectively.

A summary of the charge states found and removed are also printed to the MATLAB command window:

```
Multiple charge states found for CSN1_HUMAN (CSN) peptide KFYESKYASC (residues
332:341):
  Charge 1+ has a max stdev.p of 0.082 (intensity-weighted) across all timepoints
  Charge 2+ has a max stdev.p of 0.191 (intensity-weighted) across all timepoints
  Keeping charge with lowest stdev.p (1+) ..

Multiple charge states found for CSN1_HUMAN (CSN) peptide YLAPHVRTL (residues 357:365):
   Charge 1+ has a max stdev.p of 0.064 (intensity-weighted) across all timepoints
   Charge 2+ has a max stdev.p of 0.125 (intensity-weighted) across all timepoints
   Keeping charge with lowest stdev.p (1+) ..
```

Peptides that are removed during the data import process are displayed in the Removed Data tab table. Peptides are added to this table through two routes, firstly, unique peptides and timepoints in *State A* that are not found in *State B* and vice versa, and those that are removed via the *aggregate* function. The *aggregate* function transforms the cluster-style data into a state-like format whereby the *Uptake* and *UptakeSD* values can be easily accessed. Another feature is that the centroid mass of the peptide under *Center*, now represents the mass of the singly charged ion. For transparency and also so that errors can be spotted more easily, the replicate values for *Uptake* and *RFU* can also be found under *Uptake_Reps* and *RFU_Reps* in the table. Finally, the last column *Correction* identifies whether the Uptake data for the peptide has been back-exchange corrected.

Table 3.5. Modified '*cluster*' structure from following *Import*.

Protein	Start	End	Sequence	Modification	Fragment	MaxUptake	MHP	State	Exposure	Charge
'CSN1_HUMAN'	64	70	'ADHCPTL'	NaN	NaN	5	756.330	'CSN'	'0.250'	1
'CSN1_HUMAN'	64	70	'ADHCPTL'	NaN	NaN	5	756.330	'CSN'	'30.000'	1
'CSN1_HUMAN'	64	70	'ADHCPTL'	NaN	NaN	5	756.330	'CSN'	'5.000'	1
'CSN1_HUMAN'	64	70	'ADHCPTL'	NaN	NaN	15	756.330	'CSN'	'Sum'	1
'CSN1_HUMAN'	273	280	'ASFDHCDF'	NaN	NaN	7	941.350	'CSN'	'0.250'	1
'CSN1_HUMAN'	273	280	'ASFDHCDF'	NaN	NaN	7	941.350	'CSN'	'30.000'	1
'CSN1_HUMAN'	273	280	'ASFDHCDF'	NaN	NaN	7	941.350	'CSN'	'5.000'	1
'CSN1_HUMAN'	273	280	'ASFDHCDF'	NaN	NaN	21	941.350	'CSN'	'Sum'	1
'CSN1_HUMAN'	273	288	'ASFDHCDFPELLSPSN'	NaN	NaN	13	1778.800	'CSN'	'0.250'	2
'CSN1_HUMAN'	273	288	'ASFDHCDFPELLSPSN'	NaN	NaN	13	1778.800	'CSN'	'30.000'	2
'CSN1_HUMAN'	273	288	'ASFDHCDFPELLSPSN'	NaN	NaN	13	1778.800	'CSN'	'5.000'	2
'CSN1_HUMAN'	273	288	'ASFDHCDFPELLSPSN'	NaN	NaN	39	1778.800	'CSN'	'Sum'	2

Center	CenterSD	Uptake	Uptake_Reps	UptakeSD	RT	RTSD	RFU	RFU_Reps	RFUSD
757.600	0.032	0.905	[1×3 double]	0.0528	4.605	0.0040	18.103	[1×3 double]	1.0563
758.160	0.054	1.468	[1×3 double]	0.0684	4.604	0.0075	29.351	[1×3 double]	1.3688
757.790	0.029	1.098	[1×3 double]	0.0508	4.608	0.0014	21.965	[1×3 double]	1.0153
0.000	0.000	3.471	[1×3 double]	0.0000	0.000	0.0000	69.418	[1×3 double]	0.0000
943.020	0.013	1.227	[1×3 double]	0.0558	5.667	0.0047	17.530	[1×3 double]	0.7966
943.800	0.029	2.008	[1×3 double]	0.0613	5.670	0.0078	28.684	[1×3 double]	0.8762
943.730	0.044	1.935	[1×3 double]	0.0698	5.672	0.0065	27.640	[1×3 double]	0.9973
0.000	0.000	5.170	[1×3 double]	0.0000	0.000	0.0000	73.854	[1×3 double]	0.0000
1782.300	0.008	1.472	[1×3 double]	0.0215	6.192	0.0011	11.326	[1×3 double]	0.1655
1784.000	0.024	3.265	[1×3 double]	0.0314	6.137	0.0523	25.112	[1×3 double]	0.2412
1783.500	0.000	2.719	[2.7193]	0.0199	6.209	0.0000	20.918	[0.2092]	0.1532
0.000	0.000	7.456	[1×3 double]	0.0000	0.000	0.0000	57.356	[1×3 double]	0.0000

To the far right of the *Data Import* UI panel, a text box has been added which displays a summary of the HDX-MS data. In the latest release of Deuteros, *Summary* and *Removed Data* tabs have been added. The *Summary* tab displays a list of general information for the protein and states selected. This information includes metrics which were recommended by a recent community guideline paper[146]:

1. HDX time course
2. HDX controls
3. Back-exchange (mean / IQR)
4. Number of peptides
5. Sequence coverage
6. Average peptide length/redundancy
7. Number of replicates (and biological or technical)
8. Repeatability (average standard deviation over replicates)

The motivation for including these metrics is to increase the transparency of data used for HDX visualisation and also to shape Deuteros into a software that abides by community guidelines. Addition of this feature complexes *Objective 3: Include the latest recommendations from the HDX-MS community.* The data summary box displays metrics for State A or both State A and B when available:

```
State A: CSN (Bound)
HDX time course (min): 0.25, 1, 5, 30, 240
HDX control samples: None
Back-exchange (mean / IQR): 37.4% / 8.0%
# of Peptides: 33
Sequence coverage: 95.6%
Average peptide length / redundancy: 11.5 / 4.4
Replicates (technical): 3
Repeatability: 0.103 (average SD)

State B: CSN (Unbound)
HDX time course (min): 0.25, 1, 5, 30, 240
HDX control samples: None
Back-exchange (mean / IQR): 37.4% / 8.0%
# of Peptides: 33
Sequence coverage: 95.6%
Average peptide length / redundancy: 11.5 / 4.4
Replicates (technical): 3
Repeatability: 0.116 (average SD)
```

The protein and state name are displayed for both States A and B. A breakdown of the HDX time course found in the data and any control samples is also detailed. If back-exchange data is available and has been enabled for States A and/or B, the metric displays the mean back-exchange percentage and interquartile range (IQR). The percentage back-exchange of peptide p is calculated using the following equation (3.11):

$$Back\ exchange = \left(1 - \frac{m_{100\%} - m_{0\%}}{ND_{frac}}\right) \tag{3.11}$$

Where $m_{100\%}$ and $m_{0\%}$ are the centroid masses of the fully labelled and un-labelled peptide p, N is the theoretical maximum uptake and D_{frac} is the deuterium fraction of the labelling buffer. Both $m_{100\%}$ and $m_{0\%}$ are calculated from a set of control data that is optionally supplied to Deuteros. It should be provided as another state in the same cluster file which can be selected using the dropdown boxes beside States A and B in the data import panel. The format of control data is identical to that of State A or B but contains only references and a single timepoint that fully deuterates the peptide. The control state is processed identically to State A and B to generate state-style data. $m_{100\%}$ and $m_{0\%}$ can be found as the *Center* values for the fully labelled timepoint and reference respectively. The HDX-MS community also recognises that a degree of HDX can still occur during the quench phase[146]. However, for most state-of-the-art HDX-MS setups, $m_{0\%}$ is well approximated using the centroid mass of the unlabelled reference[146]. The back-exchange IQR is simply calculated using the inbuilt *iqr* function of MATLAB. The absolute deuterium uptake D_{abs} of peptide p can be corrected for back-exchange, essentially scaling up the uptake level to approximate the full extent of deuterium incorporation. This is performed using equation (3.12).

$$D'_{abs} = \left(\frac{D_{abs} - m_{0\%}}{m_{100\%} - m_{0\%}}\right) N \tag{3.12}$$

A few safety features have been programmed with respect to back-exchange correction with the aim of increasing awareness of how the user's data has been processed by the software. Firstly, should the user select only one state to be back-exchange corrected, an error dialog box will appear, as this disqualifies the data for differential comparisons. Next, should back-exchange not be applied to certain peptides, an error dialog box will spawn, indicating the number of non-corrected peptides. A list of these peptides can be found in the *Table* tab of the main interface under '*Non-Corrected Data*'. The value of the *Correction* column of the data table will additionally remain as '*Not corrected*' so that the peptide can be easily identified, and the user encouraged to check the input files. Finally, during plotting of any graphs, should the data table contain peptides with *Correction* equal to '*Not*

corrected', those data points will be highlighted with a yellow border to both allow instant identification and also as a cautionary note to the user to not mis-interpret the data. This feature is activated by default but can be switched off using a check box located under the *Advanced Plot Options* menu.

The *number of peptides*, *sequence coverage*, *average peptide length* and *redundancy* can be calculated through straightforward means using the data table for States A and B. For the number of replicates, Masson *et al*. suggest indicating whether a biological or technical repeat has been performed. While data from a biological repeat can be input into Deuteros by manually concatenating the cluster file while keeping all protein identifiers (such as protein name, state names and sequences) identical, this is not an intended feature of Deuteros which has been programmed to accept technically replicated data. Thus, the *replicate* metric displays the number of technical replicates that are calculated from the input data. Finally, the *repeatability* metric indicates the average standard deviation across technical replicates.

3.4.3 Data visualisation methods

Since Deuteros is a data visualisation software, these are numerous different styles of representations for both single and differential state HDX-MS. Deuteros generates eight different styles of plots: 1) coverage map, 2) redundancy map, 3) differential Woods plot, 4) single-state Woods plot, 5) multi-state Wood plot, 6) single-state butterfly plot, 7) multi-state butterfly plot and 8) the volcano plot. Each flavour of plot has been designed to suit the needs to different users and to maximise the legibility of data. Each plot also employs statistical analysis which separate peptide data into those which show significant uptake differences and those that do not. Each of the eight plot types fall into one of the three categories: 1) linear data maps (coverage and redundancy), 2) time-resolved plots (Woods and butterfly plots) and 3) ensemble plots (volcano plot). The inclusion of eight different plot and

visualisation types completes the 2D portion of *Objective 2: Develop methods of data visualisation, including 2D and 3D representations* .

3.4.4 Linear data maps: data coverage & redundancy

The coverage plot is the simplest type of plot that Deuteros generates (Figure 3.10). Although simple, the coverage map of Deuteros is easy to read from and is immediately publication quality. In addition, the side-by-side comparison of two states allow users to immediately identify regions of difference within their data. Coverage plots have only an horizontal axis of residue number and uses solid colours to represent different data types. The protein length is indicated by the strip of grey, and regions of coverage are highlighted in either of the Deuteros theme colours: 'Shocking Pink' (rgb 0.87, 0.57, 0.74) or 'Jordy Blue' (rgb 0.45, 0.67, 0.84). The protein name is shown above the plot, and above each state are the state names, percentage coverage and number of peptides for each state.

Figure 3.10. Coverage and redundancy linear maps from Deuteros. (a) Coverage map of example data indicating coverage of two states and their percentage coverage. (b) Coverage map of the common peptides between states A and B. (c) Redundancy map showing per-residue redundancy. Colour bar represents the number of redundant peptides covering a particular region. The grey bar behind each state represents the length of the protein.

In addition to the coverages of states A and B, the common coverage between the two states can also be generated. For differential comparisons, the data is pre-filtered to include peptides and timepoints which are found in both states. The common coverage map is particularly useful as a simple graphic to demonstrate the coverage between a set of experiments and identify the regions of redundancy. Next, the redundancy plot of Deuteros follows the format of the coverage plot and displays not only the coverage, but also the residue-level redundancy of each state using the 'parula' colourmap of MATLAB. The parula colour theme of MATLAB is a blue-blue/green/yellow gradient which was selected to maximise the legibility of differences between data while also increasing the clarity of the gradient for colour-blind individuals. When presenting HDX-MS, users should opt for the redundancy

plot over the coverage plot due to the extra layer of redundancy information that is available.

Linear data maps are excellent for communicating a large set of data when only peptide coverage is concerned, such as assessing the reproducibility between experiments, or when used to display a coverage summary of a multi-subunit complex. When the redundancy map style is used, these plots are particularly information rich for displaying both coverage and redundancy of proteins in a simplistic manner (Figure 3.10). We anticipate that users will want full editing control over the figures that Deuteros generates, and as such, all exported plots are saved in a vector format (e.g. scalable vector graphics, svg, or portable document file, pdf) in order to retain access to all plot elements, should editing be necessary. Vector images can be edited using a range of software, both free and paid software such as Inkscape[e] or Adobe Illustrator[f].

3.4.5 Time-resolved plots: Single and multi-state Woods plot

The Woods plot is a commonly used style of representing HDX-MS data. Where the linear coverage maps collapse the peptide data onto a single one-dimensional axis, the Woods plot adds an additional vertical axis which is typically a measure of deuterium uptake (Figure 3.11). Both absolute uptake and relative uptake are commonly plot. The Woods format also offers qualitative redundancy information since overlapping peptide regions can be easily identified. For both single and multi-state Woods formats, Deuteros applies a standard one-tailed t-test for evaluating the statistical significance of peptide uptake for each timepoint. The confidence level can be selected using the dropdown menu found under the *Advanced Plot Options*

e https://inkscape.org/

f https://www.adobe.com/uk/products/illustrator.html

subpanel. Peptides which show deuterium uptake exceeding that of the confidence interval (CI) is coloured in blue, while those which are below are deemed as showing no uptake and are in grey. At the top of the plot, is a legend bar which displays the confidence level used for statistical filtering, the numerical confidence interval in Daltons, the number of peptides in each uptake and no-uptake categories and the name of the state.

For multi-state Woods plots, two states are displayed on the horizontal axis and their corresponding deuterium uptake or relative uptake displayed on the vertical axis. Two colours are used to distinguish between peptides of the states, however there is no distinction between peptides of the no uptake category which are all plotted as grey. Equally, a legend bar is found both at the top and bottom of the plot to indicate which state has been plotted in which direction.

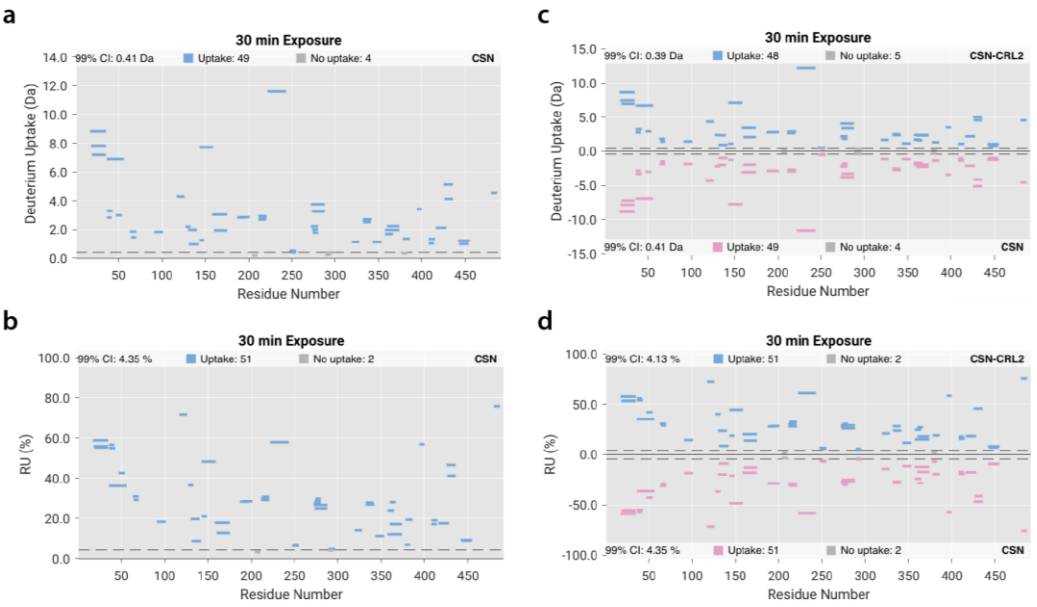


Figure 3.11. Single and multi-state Woods plots in Deuteros. (a-b) Single-state and (c-d) multi-state Woods plots of (a, c) absolute deuterium uptake in Daltons, and (b, d) the relative uptake expressed as a percentage. Each bar represents a single peptide, with the length on the horizontal axis corresponding to the peptide length. A confidence interval has been applied for statistical filtering of the peptide ensemble (dashed line). Peptides that have uptake above that of the confidence interval are shown in blue for single-state plots and blue/pink for multi-state plots. Peptides without significant uptake are shown in grey. The number of peptides in each category, confidence interval and state name are shown in the top legend bar.

3.4.6 Time-resolved plots: Differential Woods plot

Where the single and multi-state Woods plots improve on the linear data maps by adding an additional vertical axis, the differential Woods plot simplifies the multi-state Woods data by subtracting the peptide uptake of *State B* from *State A*, commonly denoted as $\Delta(StateB - StateA)$ (Figure 3.12). The differential Woods plot follows the same format as the single and multi-state Woods plots and displays a per-peptide, per-timepoint view of the comparison. Similar to other Woods plots, the differential Woods employs a confidence interval calculated using a two-tailed t-test for statistically filtering peptides. The confidence limits are centred around 0 uptake (no difference) and for timepoint measurements, are calculated according using (3.13)[140].

$$CI_t = 0 \pm \left(\frac{\sigma_t}{\sqrt{N}}\right)\alpha \tag{3.13}$$

Where σ_t is the standard deviation of the mean uptake for timepoint t, N is the number of replicates, and α is the critical value that can be found on a student t-distribution table for degrees of freedom equal to 2, e.g. α of 6.965 for 99% confidence. For sum data, where the uptake over each experimental timepoint has been aggregated to one value, the confidence limits are calculated as a simple sum of variables (3.14)[140].

$$CI_{sum} = 0 \pm \left(\frac{\sqrt{\sum_{t=1}^{n}\sigma_t^2}}{\sqrt{N_n}}\right)\alpha \tag{3.14}$$

The confidence interval of the sum CI_{sum}, is calculated as the square root of the sum of variance of all timepoints σ_t^2 divided by the square root of the number of timepoint observations N for n timepoints and multiplied by the critical value α for a particular confidence level.

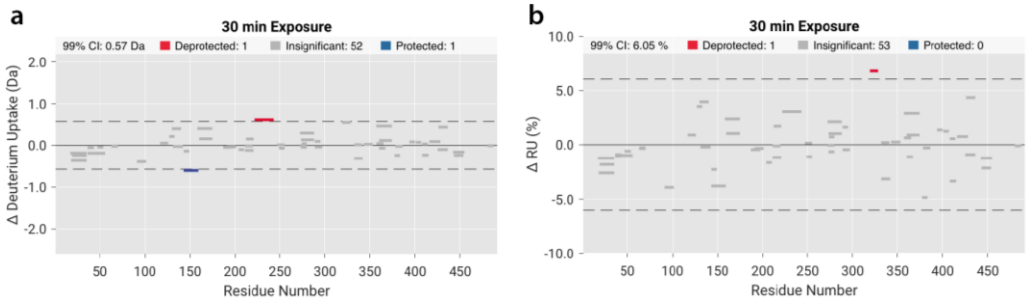


Figure 3.12. Differential Woods plot. Example of differential data plotted using (a) absolute uptake and (b) relative uptake. Each bar represents a single peptide, with the length on the horizontal axis corresponding to the peptide length. Confidence limits around 0 Da or 0 % are shown as dotted lines. Peptides with uptake > CI or uptake < -CI show significant differences between States A and B and are shown in red or blue respectively. Non-significant peptides are shown in grey. The confidence level applied, CI and number of peptides in each category are shown in the legend bar.

3.4.7 Time-resolved plots: Single and multi-state Butterfly plot

The butterfly plot is a plot style that is inherited from DynamX (Waters Corp.) due to its simplistic representation of peptide uptake data. While limited in interpretation due to the anonymised peptide identity, the butterfly plot is visually appealing and allow trends in the uptake data to be easily spotted. The butterfly plot has a similar format to the single and multi-state Woods plot in which it displays one or two states on the positive and negative vertical axis (Figure 3.13).The butterfly plot retains uptake information on the vertical axis in the form of absolute or relative uptake, however, switches the horizontal axis to an anonymised peptide index. An interpolation curve is additionally plotted to maximise the legibility trends between peptides. The utility of the interpolation curve is immediately apparent in a side-by-side comparison of the butterfly plot with and without the curve (Figure 3.14).

In addition, peptides are spread evenly along the horizontal axis, avoiding overcrowding of datapoints (as may happen in Woods plots for regions of high redundancy), however in doing so, sacrifices redundancy information. Another shortcoming of the butterfly plot format is that by standardizing the horizontal axis

to an anonymous index value, information about the peptide length is lost. Since relative uptake is normalised to the maximum theoretical uptake of the peptide and thus the peptide length, plotting relative uptake in the butterfly format is preferred over absolute uptake. This additionally leads to large differences in the shape profiles between the absolute and relative uptake plots, such as seen for the 4-5th peptides of Figure 3.13a-b.

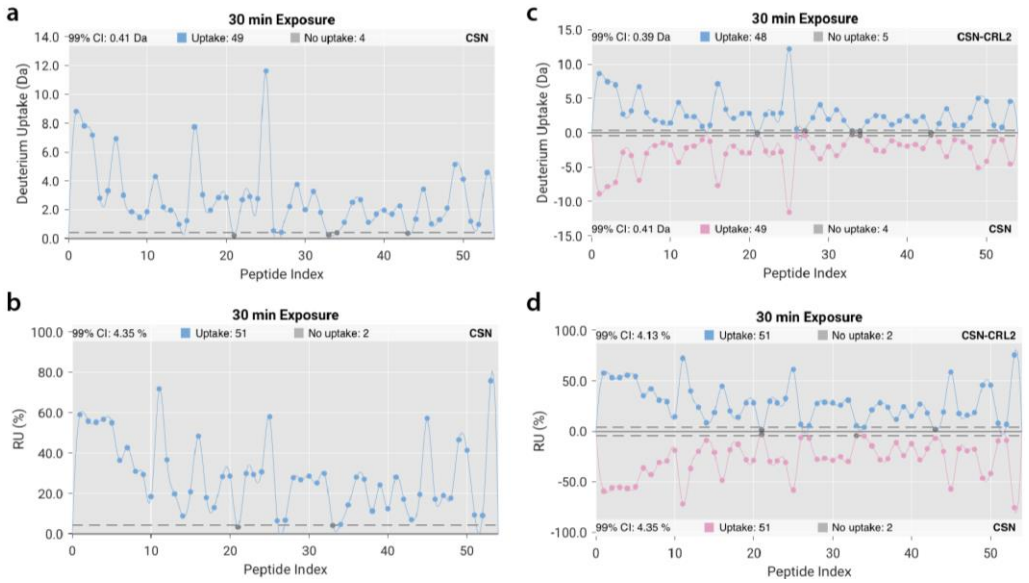


Figure 3.13. Single and multi-state butterfly plots. (a-b) Single-state and (c-d) multi-state butterfly plots of (a, c) absolute deuterium uptake in Daltons, and (b, d) the relative uptake expressed as a percentage. Each scatter point represents a single peptide. A confidence interval has been applied for statistical filtering of the peptide ensemble (dashed line). Peptides that have uptake above that of the confidence interval are shown in blue for single-state plots and blue/pink for multi-state plots. Peptides without significant uptake are shown in grey. The number of peptides in each category, confidence interval and state name are shown in the top legend bar.

Since a feature of the butterfly plot is that peptides similar in sequence position are found next to one another, it is imperative that the underlying peptide data for the plot must first be sorted via the start residue number, such that the peptide index appears in the correct order. Correctly ordering peptides allows the 'smoothness' of the data to be assessed as an additional feature of the plot. The 'smoothness' refers to how dramatically the deuterium uptake of a region of sequence changes, since a group of peptides from the same stretch of sequence should in theory exhibit similar

uptake levels. Assuming that there are no lapses in the coverage of the protein and that each sequential peptide is of similar length and shifts by a uniform number of residues, the interpolated curve should closely match that of a residue-level deuterium uptake plot.

Statistical filtering for butterfly plots are performed as one-tailed t-tests in the same manner as for the differential Woods format. When applied to a butterfly plot, the confidence interval is used as a threshold to distinguish between peptides that show deuterium uptake and no uptake. For example, on a relative uptake plot, the percentage uptake of a peptide must exceed that of the confidence interval for the peptide to be classified as one which has uptaken deuterium. Practically, this filtering can be used to quickly distinguish between regions of the protein which may be solvent exposed or buried. We envision that this feature will be useful for membrane proteins or large soluble complexes which harbour extensive buried regions.

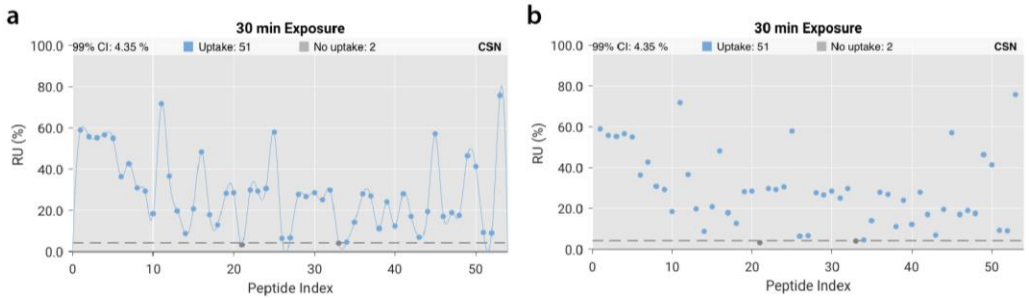


Figure 3.14. Utility of the interpolation curve in butterfly plots. The same example data is plotted (a) with and (b) without interpolation curves in the butterfly plot format.

3.4.8 Ensemble plot: The Volcano plot

The volcano plot is an ensemble plot type whereby all datapoints of all timepoints are simultaneously shown in a single plot for a differential comparison, typically performed as $\Delta(State_B - State_A)$. It takes its name from the appearance of the graph which when plotted, resembles that of an erupting volcano. The volcano plot plots a measure of uptake change between two states against the p-value of its

replicate measurements (Figure 3.15). Thus, a volcano plot assesses both the physical measured change in deuterium uptake between States A and B, and the variation of the measurement. Typically, the fold-change is plot against the p-value of the data however in Deuteros, both fold-change and ΔMass can be used as a measure of change between peptides of State A and B (Figure 3.15a-b). To exacerbate the appearance of the volcano, $log_2(Fold-change)$ and $-log_{10}(p-value)$ is typically plot. The p-value is taken to the $-log10$ in order to visually enhance differences between datapoints and reverses the vertical axis such that lower p-values lead to the datapoint appearing at the top of the graph.

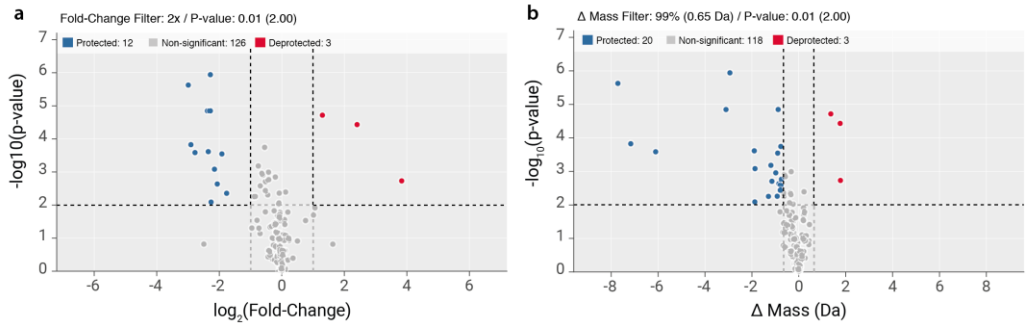


Figure 3.15. Volcano plots of Deuteros. (a) Fold-change and (b) absolute change in mass of peptides between State A and B. Dashed vertical lines represent the fold-change (e.g. 2, 3 or 4-fold) for (a) and confidence limits for (b) for the selected confidence level. Dashed horizontal line represents the p-value across technical replicates. Peptides with fold-change or Δmass greater than positive vertical bar and above the horizontal p-value threshold are in the deprotected category (red). Those with fold-change or Δmass less than the negative vertical bar and above the p-value threshold are in the protected category (blue). Those not in deprotected or protected categories show non-significant changes in the $\boldsymbol{\Delta(State_B - State_A)}$ comparison. The numerical values for each horizontal and vertical filter and number of peptides in each category are shown in the legend bar.

Figure 3.16. Adding interactivity to volcano plots. Peptide data in volcano plots can be interactively accessed via clicking on individual datapoints. Peptide sequence information,

timepoint, the Δ Mass (including absolute deuterium uptake for states A and B), the P-value and number of replicates can be shown for each peptide. Data tips remain on exported plots.

The volcano plot is the most statistically rigorous plot in Deuteros 2.0 since it employs two tests to identify peptides which show significant differences. The two statistical tests set up bidirectional filters and are used to determine whether a peptide is in one of the deprotected, protected or non-significant categories similar to the differential Woods plot. For the fold-change-type volcano (Figure 3.15a), the x-direction filter can be a positive or negative 2, 3 or 4-fold change in deuterium uptake. The p-value filter is calculated as the $-log_{10}(p-value)$ of the p-value selected by the user. For Δmass-type volcano plots (Figure 3.15b), the vertical filter is a confidence interval calculated at a particular confidence level[140]. Selecting a confidence level for the ΔMass-type plot changes both the x and y-direction filters while it only affects the y-direction filter for the fold-change-type volcano plot. The data underlying each peptide, including as the peptide start and end residues, sequence and exposure, can be accessed interactively via clicking on individual datapoints within the volcano plot (Figure 3.16).

3.4.9 Structural visualisation

Another major feature and utility of Deuteros is its ability to transform HDX-MS data into formatting scripts that can be applied onto structures of proteins in order to visualise data more easily. The features described in this section competes the 3D portion of *Objective 2: Develop methods of data visualisation, including 2D and 3D representations* .This function is carried out by the '*Export to Molecular Graphics*' subpanel of the Deuteros GUI. Currently exporting to the *PyMOL*[g] and *Chimera*[h] molecular graphics viewers are supported. This feature supports the exporting of eight different data types: 1) coverage, 2) redundancy, 3) differential Woods, 4)

g https://pymol.org/2/

h https://www.rbvi.ucsf.edu/chimera/

single-state Woods, 5) multi-state Woods, 6) single-state butterfly, 7) multi-state butterfly and 8) volcano (Figure 3.17).

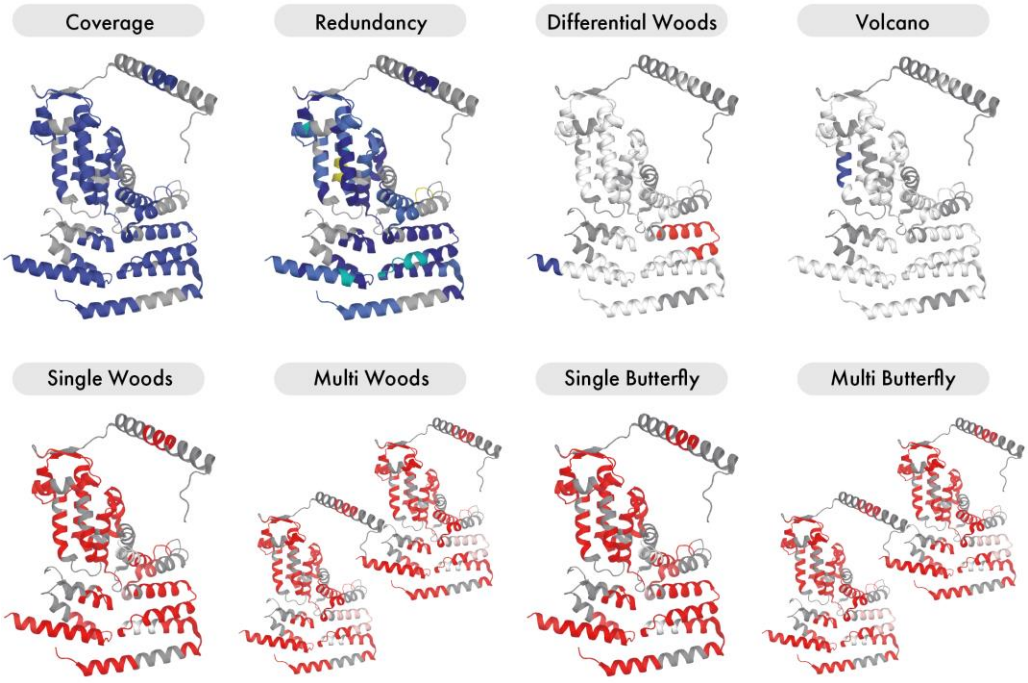


Figure 3.17. Summary of structural formatting from Deuteros 2.0. Export styles of (a) PyMOL and (b) Chimera for eight plot types. The colours used in each plot type correspond to colours used in each of the two-dimensional plots in Deuteros.

To project HDX-MS data onto molecular structures, the data is first plot using the plot type of interest in Deuteros. The user then selects the relevant export options in the '*Export to Molecular Graphics*' subpanel. Within this subpanel, several export options are available for customisation. The '*Export to*' dropdown menu allows the selection between *PyMOL* and *Chimera* graphical viewers (Figure 3.18). The PDB chain ID of the protein structure to apply HDX-MS data to, is required to be input into the '*Chain*' text box. Deuteros allows chain IDs to be input in two formats: either as a single chain (i.e. 'A') or multiple chains separated by commas (i.e. 'A,B,C'). An error dialog window spawns in the case that the user supplies a non-string or empty chain ID. Next, the '*Export Coverage data*' and '*Export Advanced data*' check boxes allow

the user to select which currently plotted data is to be exported. Under the '*Export Coverage data*' section, colour options for areas of coverage, no coverage and the redundancy colour map can be selected. Finally, the '*Export*' button can be pressed to bring up a save dialog window which allows the user to select the filename and location that the exported formatting scripts are saved under.

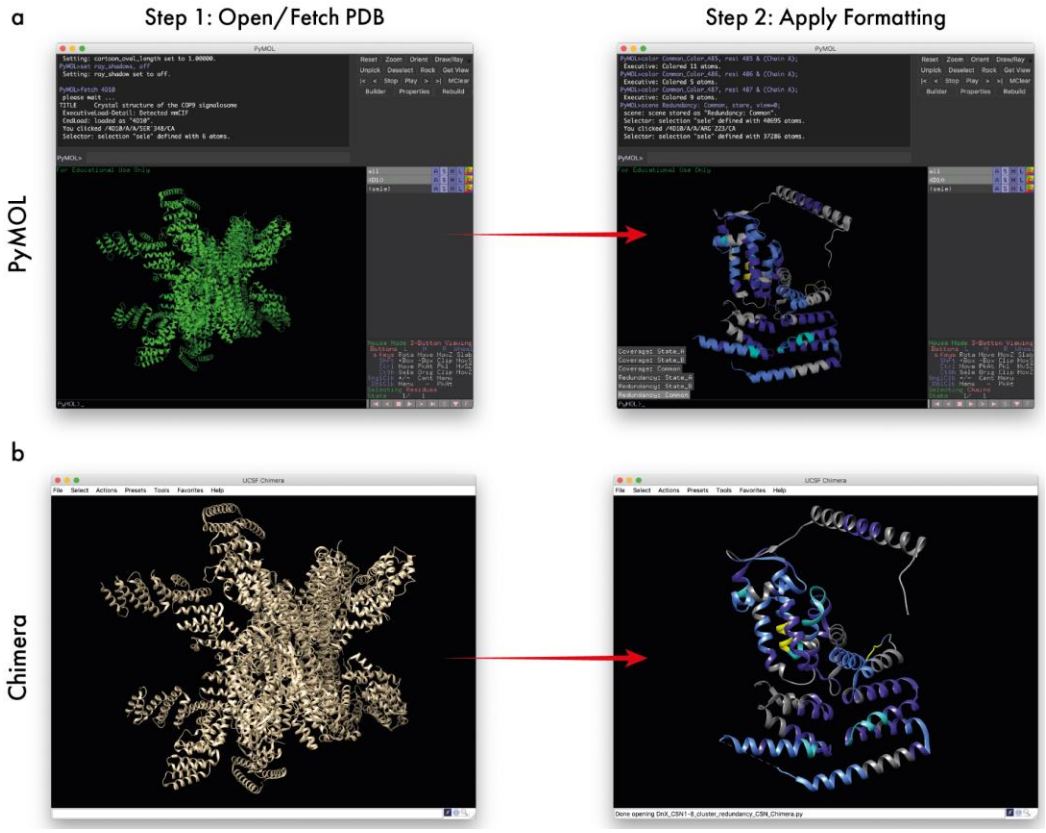


Figure 3.18. HDX-MS data can be projected onto molecular structures in (a) *PyMOL* and (b) *Chimera* in two steps. The first step requires the model to be either opened locally or fetched from the PDB database using the inbuilt functions of *PyMOL* or *Chimera*. The second step applies the HDX-MS colour formatting using a customised python script. For *PyMOL*, a *.pml* script can be drag-and-dropped into the *PyMOL* window to apply colouring. In *Chimera*, a *.py* script is opened via the *File* > *Open* drop-down menu.

When interpreting data on 3D structures, two aspects of the model must be considered. Firstly, it is important for the user to be aware of any residues and regions that might be missing from the model. For obvious reasons, data cannot be plotted onto missing regions of the model, and as such, users may unintentionally ignore biologically interesting areas of the data. Residues may be absent from the

structure due to gaps in the experimental data used to generate the model. The reasoning behind why there may be missing regions from a structure, is technique dependent. For example, for structures determined using X-ray crystallography, the highly flexible or disordered regions of a protein structure may manifest as missing residues in a PDB, due to a lack of occupancy in the crystal lattice for that region. Therefore, it is suggested that users should first familiarise themselves with the models that they wish to use for data visualisation. To aid in the transparency of missing residues, Deuteros 2.0 has been programmed to print a notice for the user should residues be absent when a formatting script is applied to a structure. Missing residues are calculated and printed on screen for only the user defined chain ID. The pseudocode for the calculation is shown below:

```
Pseudocode:
make empty list to store residues called "residues"
iterate over all coordinates of user specified chain, append residue numbers to "residues"
make a full residue list called "complete", by taking first and last values in "residues" and include all in between integers
generate "missing_residues" list by comparing "residues" against "complete"

if length("missing_residues") is greater than 0:
  print missing residues found

Original code:
residues = []
iterate_state 1, (chain A & n. CA), residues.append(int(resi))

original_list = [x for x in range(residues[0], residues[-1] + 1)]
num_list = set(residues)
missing_residues = list(num_list ^ set(original_list))

if len(missing_residues) > 0:
  print "NOTE: %s Missing residues were found: %s" % (len(missing_residues), ', '.join(str(x) for x in missing_residues))
```

The second aspect to consider during 3D data visualisation is whether or not the residue numbers in the PDB matches that of the HDX-MS data. The sequence and peptide start and end values from HDX-MS are sourced from a FASTA file that is initially supplied to PLGS. Typically, protein sequences within the FASTA sequence file are accessed from the UniProt[i] database. We suggest that protein sequences in the PDB file should be checked and matched with the HDX-MS data such that the residue numbers are aligned appropriately. Failure to do this will result in an offset to the colouration of the data applied to the structure and may be detrimental to the interpretation of the data. While this issue could be avoided by taking advantage of and using the sequence of the peptide (e.g. KKDVGLHPS) to identify the correct region of the protein in *PyMOL* or Chimera, this method has several shortcomings. Firstly, it is not uncommon for structures solved via X-ray crystallography, to possess mutations, deletions or other gene modifications, which would result in differences between the PDB and HDX-MS sequences. Secondly, HDX-MS attracts many systems which are difficult to characterise due to their structural elusiveness or high degree of dynamics, such as highly flexible, membranal or disordered proteins. These systems may in turn lack crystallographic representation, leading to homologous structures being used in their stead, and again, a difference between PDB and HDX-MS sequences. Finally, the colouration of a peptide may be missed entirely if even a single residue is missing from the PDB. For example, if the KKDVGLHPS peptide was found in Deuteros 2.0 to be deprotected, but only GLHPS is present in the PDB model, no colouration will be applied since the full sequence cannot be located. In summary, a degree of caution must be exercised by users to ensure that the HDX-MS data is appropriately visualised on 3D models. We suggest two aspects of the structure should be checked as a minimum: 1) whether residues are absent from the structure and 2) whether the protein sequence of the PDB is in the correct number frame as the HDX-MS data.

---

i https://www.uniprot.org/

3.4.10 Formatting in PyMOL and Chimera

Formatting scripts generated for *PyMOL* and *Chimera* take advantage of the software's ability to parse and run python-based scripts. Although the interfaces and formatting scripts of *PyMOL* and *Chimera* are vastly different, HDX-MS data can be applied onto molecular structures in just two simple steps, beginning from the opened software window (Figure 3.18). The first step requires a model of the protein of interest to be opened. This can be done either by accessing the file in a local directory, or alternatively can be 'fetched' via the PDB database if an active internet connection is available. In *PyMOL*, fetching a PDB file can be easily performed by inputting directly into the command-line '*fetch 4D10*' to import the 4D10 PDB model. In *Chimera*, fetching is done through the *File* > *Fetch* by ID dropdown menu and inputting the PDB ID.

Once a PDB has been loaded into the software, the formatting script can be applied in the next step. In the most recent major update of *PyMOL* version 2.0 onwards, *drag-and-drop* is supported for opening files, and as such the formatting script can be dragged onto the window directly to apply HDX-MS colouration. In *Chimera*, this feature is not available and instead requires the under to open the formatting script through the traditional *File* > *Open* menu. For *PyMOL*, the formatting script utilises the file extension '*.pml*' while it is a '*.py*' file for *Chimera*. Both scripts are python based, however the formatting syntax is different and as such require different functions included in Deuteros to parse HDX-MS data into software-specific commands. Due to differences between software and plot types, the formatting scripts generated from Deuteros differ greatly however largely follow the format shown in (Figure 3.19).

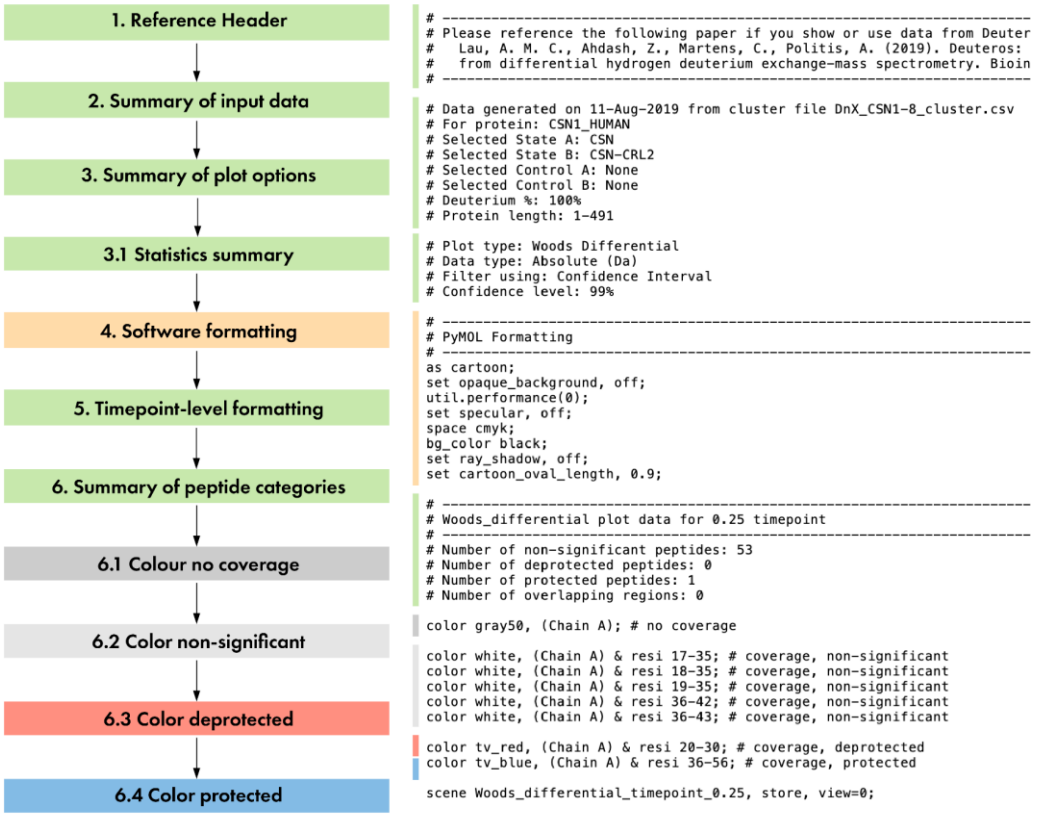


Figure 3.19. Layout of Deuteros formatting script. Left shows a flowchart representation of the formatting script layout for most of Deuteros' plot types. Green sections are those that are identical across all plot types. Orange are software-specific sections. Grey, light grey, red and blue are plot type and timepoint specific sections. An example script from the CSN1 script designed for *PyMOL* is shown on the right.

The formatting script contains a list of *PyMOL* or *Chimera* commands which transposes the visualised data from 2D plots onto 3D structures. To do this, the peptide categories of each 2D plot, are converted into colour commands which apply a colour to the region of the peptide on the 3D structure. For example, visualising a differential Woods plot, applies a red/white/blue and grey colour scheme to the 3D model to mark regions of deprotection, no significant difference, protection and no coverage. The software-specific syntax is shown below:

PyMOL:

```
# hide all cartoon representations except for chain A
hide cartoon, not chain A

# color all residues of chain A, in gray
color gray, (Chain A)

# color residues 17-35 of chain A, in white
color white, (Chain A) & resi 17-35

# color residues 20-30 of chain A OR chain B in red
color red, (Chain A | Chain B) & resi 20-30
```

Chimera:

```
# hide all ribbon representations except for chain A
runCommand("~ribbon ~ :.A")

# color all residues of chain A, in gray
runCommand("color gray :.A")

# color residues 17-35 of chain A, in white
runCommand("color white :17-35.A")

# color residues 20-30 of chain A and residues 20-30 of chain B in red
runCommand("color red :20-30.A,20-30.B")
```

While both *PyMOL* and *Chimera* are supported by *Deuteros 2.0*, it is important to point out that *PyMOL* is inherently better suited for displaying HDX-MS data due to the software's '*scene*' feature which allows interactive buttons to be programmed into the bottom left of the *PyMOL* GUI. Scene buttons allow users to toggle between different data representations, such as between coverage and redundancy, or between different timepoints of the data, without having to load a new file or spawn a new *PyMOL* session. As such, a direct benefit of coding the formatting script for

PyMOL, is that potentially, a single *.pml* file can store all data formatting types. Another software feature that makes *PyMOL* advantageous to *Chimera*, is '*grid*' view, which allows two or more structures or states to be viewed simultaneously. This feature is especially useful for side-by-side comparisons of two HDX-MS states, such as for multi-state Woods or butterfly visualisation. Structures presented in the grid, also benefit from scene buttons, providing another layer of customisable data representation for the user.

In contrast, Chimera does not possess any features equivalent to scene buttons or grid view, reducing the efficiency of data visualisation. The lack of scene buttons for toggling between data types or timepoints, means that these formatting outputs need to be separated into different formatting scripts. Coverage and redundancy data in Deuteros 2.0 is generated for *State A*, *State B* and *Common* peptides between the two states, giving a combination of 2×3 representation types. Where *PyMOL* can display these 6 data types in a single formatting script using scene buttons, 6 separate files are output for *Chimera*.

3.4.11 Structural projection of data onto molecular structures

Mapping the coverage and redundancy data onto molecular structures allows the generation of informative but simplistic figures which depict which regions of the protein are covered by the HDX-MS experiment and to what extent. For specialised systems such as integral membrane proteins, or those with extensive buried surface area, mapping the regions of coverage onto structures is especially useful for demonstrating areas of high and low coverage. The ability to visualise the coverage of the experiment in the context of the biological structure, can also be helpful in experimental design, whereby different proteolytic enzymes can be tested for their efficacy in accessing certain regions of interest in the system.

Both coverage and redundancy information for both single and multi-state HDX-MS datasets can be quickly visualised on 3D structures using the coverage and

redundancy formatting scripts. To generate residue-level values that can then be used for colouring molecular structures, the state data table of peptides used to generate the coverage and redundancy plots, are formatted into a residue-level list along with the number of times the residue appears over the whole peptide ensemble. In *PyMOL*, scene buttons are added to allow users to toggle between the coverage and redundancy of each experimental state, and the set of common peptides between the two states. A myriad of colouring options are available to select from within the '*Export to Molecular Graphics*' subpanel of *Deuteros 2.0*. It is also important to note that no colour bar (as can be seen in Figure 3.20) is generated by default or shown in the *PyMOL* or *Chimera* session. The colour bar of appropriate scale can be generated and exported from the redundancy map plot of *Deuteros 2.0*.

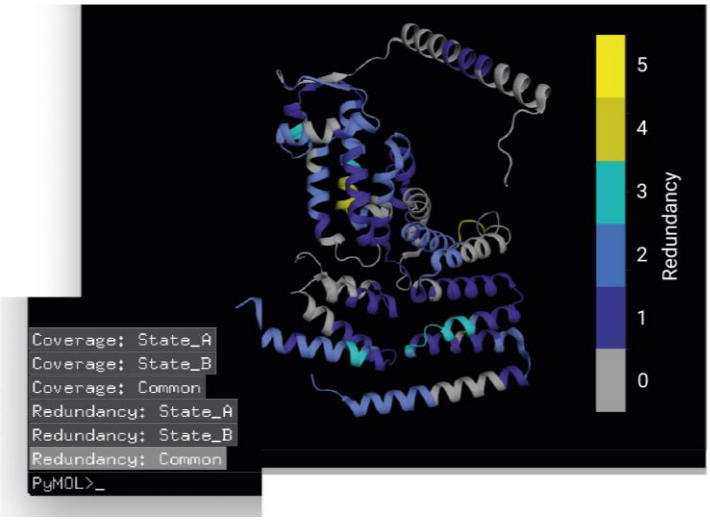


Figure 3.20. Visualising redundancy on 3D models using PyMOL. Example redundancy data shown for the CSN1 complex. Colour bar represents the residue-level redundancy. The insert displays the set of scene buttons that can be used to toggle between data types.

Next, the per-residue deuterium uptake can be visualised by exporting formatting scripts from either the butterfly or non-differential Woods plots (Figure 3.21). To generate the residue-level data, the uptake level at each residue number is approximated by averaging the uptake over all peptides. This also means that the resolution and accuracy of the residue-level uptake is dependent on the redundancy of the residue, and also the lengths of the peptides that were used to calculate the

value. Data generated using this method is typically not seen or presented as true "residue level" uptake but is only necessary for converting a table of peptides into a linear data list that can then be applied to each residue of a PDB. Given that both the resolution and accuracy of the uptake values depend heavily on redundancy, it is suggested that users should review their uptake plots alongside the redundancy plot to assess the relative confidence of certain residues.

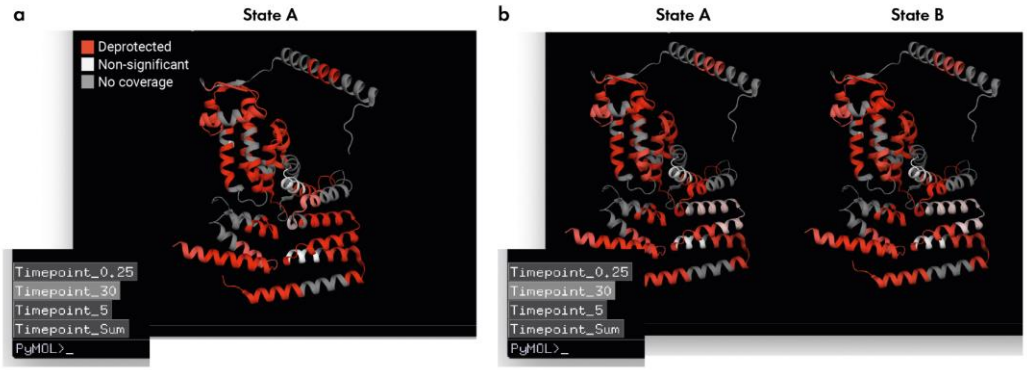


Figure 3.21. Visualising (a) single and (b) multi-state uptake data on 3D models in PyMOL. Regions of the protein without experimental coverage are coloured in grey, Areas showing non-significant uptake are in white and those with significant uptake are in red. Scene buttons allow data for each timepoint to be toggled between.

There are four export types in Deuteros 2.0 which project deuterium uptake onto structures: single-state and multi-state butterfly or Woods visualisation styles. In reality, there is repetition in the data that is presented by these four styles. Single-state butterfly and single-state Woods plots show the same data and employ the same confidence filtering, resulting in identical visualisations when formatting scripts are applied to structures. The same State A data is additionally used for the multi-state butterfly and multi-state Woods plots. Both single-state and multi-state Woods and butterfly plots utilise scene buttons for toggling between timepoint data (Figure 3.21). As mentioned previously, the 'grid view' feature of PyMOL is used for multi-state formats in order to provide side-by-side comparisons of two states. In practice, if there are few regions of significant uptake difference between States A and B, it may be difficult spot these differences by eye in the graphical viewer.

Instead, regions of significant uptake differences are better pinpointed by using one of the differential Woods or volcano plotting styles.

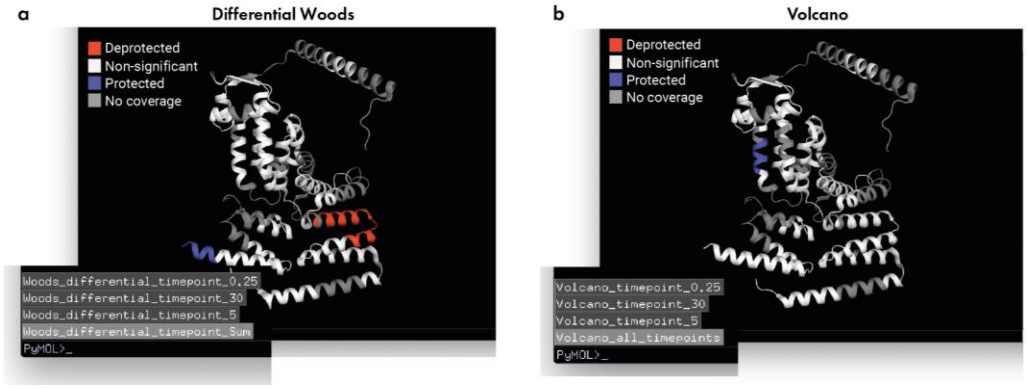


Figure 3.22. Visualising differential HDX-MS data using (a) differential Woods filtering and (b) volcano filtering methods. Regions without experimental coverage are coloured in grey. Significantly protected regions are in blue, and deprotected regions are in red. Areas without significant differences are in white. Scene buttons allow toggling between experimental timepoints.

The differential Woods or Volcano plot simplifies a comparison between two experimental states by reducing the information content to only the difference between the two states (Figure 3.22). To generate the residue-level series for differential Woods and volcano formats, the same peptide-level data table used to plot the 2D graphs are re-used. It does not calculate the residue-level uptake map of each state and then compare between them. Converting to residue-level first, would bias the comparison if there are significant differences between the redundancy of the data between the differential states. By using peptide-level data, only peptides that are found in both states are compared and used to determine whether a significant difference is measured. In both differential Woods and volcano plots, peptide categories are shown for each individual timepoint and accessed via scene buttons in *PyMOL*. For differential Woods, the additional 'sum timepoint' is also available to provide consistency with the 2D differential Woods plot. The reader may have noticed however, that for the volcano style visualisation, the sum series has been replaced by 'all timepoints'. This scene instead shows all regions which have been identified as significant in any of the constituent timepoints. The inclusion of

this 'all timepoint' series as an alternative, was due to the fact that sum data could not be simply filtered in a volcano plot, using the same confidence interval and p-value filters calculated from the constituent timepoints that compose the series in the first place.

A comparison of data filtered using differential Woods and volcano plots reveals subtle differences between the two methods. As expected, applying an additional p-value filter in the volcano plot, results in fewer peptides identified as statistically significant. A p-value filter essentially removes datapoints such that only those with the greatest evidence against the null hypothesis[j] remain. This is visible when visually comparing structural models of the 30 min timepoint in Figure 3.23. As mentioned earlier, a potential issue arises when regions of the model are missing from the PDB file. These peptides have been highlighted in Figure 3.23a and the colouration is entirely missing from the structures presented in Figure 3.23b-c.

In summary, this chapter has demonstrated the myriad of structural visualisation styles that Deuteros 2.0 produces. Both single-state, multi-state and differential comparisons are available, and for both *PyMOL* and *Chimera* visualisation software. The ability for users to transpose 2D HDX-MS data directly onto 3D structures in two simple steps, provides greater access for meaningful interpretation of results. Visualisation of coverage and redundancy information as mentioned, provides several benefits, one being that the visualisation can help with experimental design, such as optimising sample preparation steps and acquisition parameters to maximise the number of useful peptides recovered from an experiment. Visualising deuterium uptake levels using single-state and multi-state representations allow uses to review which regions of their protein are responsive to HDX and can inform on exposed and buried areas. Finally, the ability to project differential HDX-MS onto structures, is the hallmark of *Deuteros 2.0*. Using volcano plots for stringent filtering

j Null hypothesis/H0 : No difference in uptake between States A and B

of potentially interesting peptides, allows regions of biological interest to be pinpointed with greater confidence.

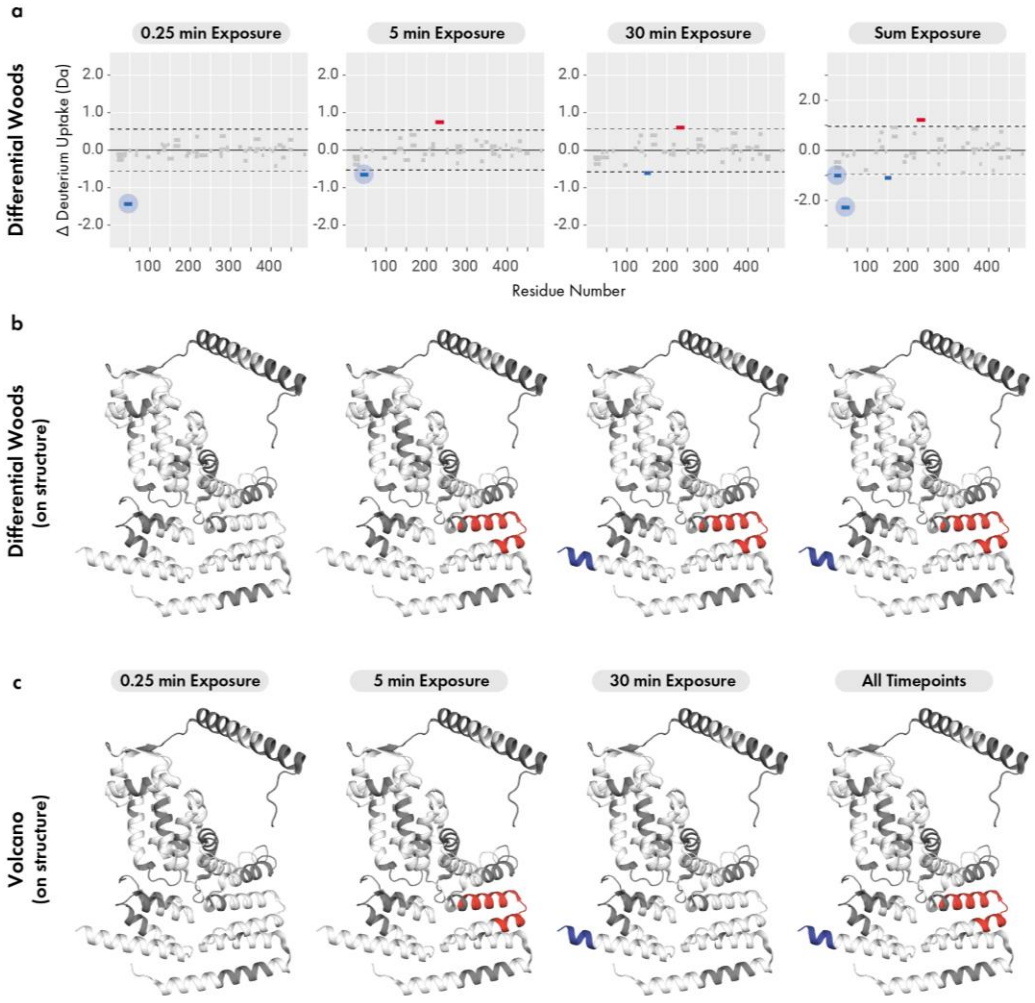


Figure 3.23. Comparison of data filtered using differential Woods and volcano plot formats. (a) 2D differential Woods plot for example data over 0.25, 5, 30 min and sum timepoints. Each bar represents a single peptide. The 99% confidence intervals are shown by the dotted lines and is used to statistically filter peptides to classify them into one of three categories: $\boldsymbol{\Delta DU > CI}$ are deprotected (red), $\boldsymbol{\Delta DU < -CI}$ are protected (blue) and those with $\boldsymbol{-CI < \Delta DU < CI}$ are non-significant (grey). Peptides which are missing from the PDB model in (b-c) are highlighted by blue circles. (b) Structural representation of differential Woods data shown in (a). Cartoon colours correspond to peptides categories determined in (a). Grey and white indicates peptides without coverage, and those that show non-significant $\boldsymbol{\Delta DU}$ respectively. (c) Structural representation of volcano plot data (Figure 3.20). Peptide colours follow that of (b).

3.5 Discussion

In section 3.2 Aims & Objectives, we described five aims of Deuteros 2.0:

1. To provide an easy to use platform for statistical analysis and visualisation of HDX-MS data
2. To provide a method of extracting meaningful information from HDX-MS data in the context of structural interpretations
3. To develop a foundation software that can be dynamically adjusted to the needs of the evolving HDX-MS field
4. To keep user interaction to a minimum where unnecessary
5. Free to use, offline and accessible on most computers

In software engineering, the quality of software can be evaluated in terms of functional and structural quality. Functional quality refers to the ability of the software to fulfil functional requirements, both in terms of design and software features. On the other hand, structural quality refers to the robustness and structure of the underlying programming. With this in mind, we can simplify the five above aims of Deuteros 2.0 into those that can be assessed as functional or structural qualities. Aims 1, 2, 4 and 5 relate to software and data analysis features and as such, are functional qualities. The accessibility of the software is another important functional quality to discuss. This encompasses considerations such as whether the software is easy to physically access (e.g. webserver, standalone installation or command-line), the skill level required (is technical experience needed?) or how easy input data can be generated, especially if dependent on upstream software. Aim 3 on the other hand, refers to the development of the back-end code into an adjustable format that can be easily expanded. Each of the aims in the context of functional and structural quality will be discussed and compared with other software in the statistical analysis and data visualisation category.

3.5.1 A comparison of Deuteros 2.0 with other statistical and visualisation software

As pointed out earlier, several software were developed for statistical analysis and visualisation of HDX-MS data. These are MSTools, MEMHDX, HDX-Analyzer and HDX-Viewer. To systematically review the functional quality of Deuteros 2.0 against each of its competitor software, we assembled a list of features provided across the MSTools, MEMHDX, HDX-Analyzer, HDX-Viewer and Deuteros 2.0 software and generated a comparison table to better summarise similarities and differences (Table 3.6). The following sections will compare the accessibility of the software to users, difficulty of generating the input data and each of the software features to those of Deuteros 2.0.

Table 3.6. Comparison of statistical and visualisation software

Parameters	MSTools	MEMHDX	HDX-Analyzer	HDX-Viewer	Deuteros 2.0
Year of release	2010	2016	2011	2019	2019
Software accessibility					
Implementation	Webserver	Webserver	Standalone	Webserver	Standalone
Free?	Yes	Yes	Yes	Yes	Yes
Available	Yes	Yes	No	Yes	Yes
Language	PHP, JavaScript	R	Python, R	HTML5, JavaScript	MATLAB
Is installation dependent on software?	No	No	No	No	Yes, MATLAB (commercial) or MATLAB Runtime Library (free)
Input format	Plain text (.txt)	DynamX Cluster (.csv)	Excel (.xls)	DynamX (.pml) or HDExaminer (.csv)	DynamX Cluster (.csv)
Editing needed?	Yes	Yes	Unknown	No	No
Software Features					
Statistics	None	Mixed effects model, Chi-Squared Test	Multiple regression, paired t-test, ANCOVA	No	Confidence interval, unpaired t-test
Differential uptake	No	Yes	Yes	No	Yes
Single-state uptake	Yes	No	No	Yes	Yes
Multi-state uptake	Yes	No	Yes	No	Yes
Kinetics plot	No	Yes	Yes	No	No
Theoretical HDX analysis	Yes	No	No	No	No
Coverage/Redundancy	Yes	No	No	No	Yes
Structural visualisation	Yes	Yes	Yes	Yes	Yes
Export figures	Yes	Yes	Unknown	No	Yes
Export data tables	Yes	Yes	Unknown	No	Yes
Export structures	Yes	No	Unknown	Yes	Yes

3.5.2 Software accessibility

We first reviewed each software in the statistics and data visualisation category in terms of software accessibility and implementation. The more popular implementation method of HDX-MS software is in the form of webservers (MSTools, MEMHDX and HDX-Viewer), while only HDX-Analyzer and Deuteros are standalone software. While webservers are simpler to access than standalone installations from the user standpoint, webservers have several shortcomings. Firstly, they are more prone to downtime, both scheduled (maintenance and upgrades) and unscheduled (server crashes). Secondly, webservers are more difficult to maintain for long periods of time since they require someone capable of maintaining them, compared to a single installation of a software that can be used indefinitely. Finally, webservers are susceptible to web attacks which in combination with requiring users to upload data to a remote server, may compromise data. While webservers are easier to access, providing that standalone software such as Deuteros 2.0, does not require difficult installation, software licenses or other software dependences, distributing software as standalone installations does not restrict access to data analysis facilities. Out of the five software compared, we were able to access all with the exception of HDX-Analyzer which appeared to be no longer available from the original download link. This highlights an issue in the HDX-MS software field in which some useful tools may become unmaintained and eventually disbanded. Deuteros 2.0 is only software written in a programming language (MATLAB) that is dependent on an external software license for editing its source code. For the average user however, there is the option of installing the MATLAB Runtime library which is freely available from MathWorks, however this means compared to the other webservers (MSTools, MEMHDX, HDX-Viewer), Deuteros 2.0 requires additional installation steps and is the least accessible in this category.

3.5.3 Accessibility of input data

The five software utilise a myriad of different methods for interpreting data from an input file that is typically produced from data processing software such as DynamX for Deuteros 2.0 and MEMHDX. Since there are numerous software performing the data processing step and these can be in turn restricted by instrumentation, each software has evolved from different necessities and typically do not utilise the same input files. By far the most comparable software to Deuteros 2.0 in terms of data input, is MEMHDX which takes the cluster output from DynamX. Although both software are designed to be used in conjunction with DynamX, there are differences in the handling of the cluster file. In Deuteros 2.0, reading the peptide data from the DynamX cluster file is performed in a single step that requires no additional manual work. On the other hand, the same cluster file to be analysed by MEMHDX, requires first a significantly laborious editing step that involves manually adding the replicate number for each peptide datapoint. The time taken for an average user to add this information to the MEMHDX input file will vary depending on the experimental complexity, data quality and number of peptides, but may be between 5-20 minutes for each cluster file. For studies consisting of many experimental states, proteins or replicates, the requirement of manually adding to the cluster file, makes MEMHDX significantly less accessible to use for data analysis than Deuteros 2.0. From a programmatic standpoint, repetitive tasks such as simple data labelling, should be performed by the software and not the user. As such, Deuteros 2.0 provides better accessibility since the time taken to connect between DynamX and Deuteros 2.0 is effectively non-existent.

We observed an interesting correlation which was that recent software such as MEMHDX, Deuteros 2.0 and HDX-Viewer (2016, 2019 and 2019 respectively) were more reliant on upstream data processing software to output data in a standardized format (such as the cluster format from DynamX), rather than earlier software (MSTools, 2010 and HDX-Analyzer, 2011) which utilised simple text files which were

vendor and software-independent. A good example of this is with the 'Draw Map' module of MSTools which generates coverage and heat maps for HDX-MS states. Three files can be supplied to 'Draw Map': 1) a sequence file that contains a simple list of peptides, 2) a digest file that contains % deuterium values for each sequence at various timepoints and experimental conditions, and 3) a structure file that includes annotations of secondary structure and domain classification for the peptide list. Generating these lists in the correct format for MSTools is significantly time consuming compared to more modern data formats such as the cluster file. An increase in the complexity and size of systems being studied by HDX-MS has also meant that modern peptide data files are comparatively larger than those of the last decade. The relationship between software age and dependence on upstream processing data formats is also visible in the most recent addition to the HDX-MS software family - the HDX-Viewer webserver. HDX-Viewer provides two options of source data - DynamX or HDExaminer. Interestingly however, unlike MEMHDX or Deuteros 2.0, HDX-Viewer does not make use of the DynamX cluster file, or any of the tabulated formats (state or difference). Instead, HDX-Viewer requires the PyMOL formatting script that DynamX generates along with a PDB model that uptake data is projected onto. The PyMOL *.pml* formatting file is similar to that generated by Deuteros 2.0 and has been designed to conduit peptide data in DynamX to protein models in PyMOL. In terms of input data accessibility, the PyMOL *.pml* file from DynamX is easy to generate and does not require any additional modifications before inputting to HDX-Viewer. The alternative input is a csv file generated from the commercial software HDExaminer which includes the columns for example: *Start, End, 5s, 10s, 5s - spread, 10 - spread*, where the spread values are a measure of replicate variance. Since HDX-Viewer provides two methods of importing data into its web-based graphical viewer, this increases the accessibility of the software to many additional users.

3.5.4 Comparison of statistical methods

A review of the capabilities of each software (Table 3.6), reveal that the majority are developed for statistical analysis for differential HDX-MS and not single-state applications. Although MSTools has features for differential HDX-MS, it does not employ statistical tests of any kind for significance testing. HDX-Viewer does not perform differential nor statistical analysis. Software with features for performing statistical analysis of HDX-MS data and downstream visualisation include HDX-Analyzer, MEMHDX and Deuteros 2.0, with each of these employing different statistical methods. Among these three software, HDX-Analyzer is unique in that statistics can be applied to either the deuterium uptake or centroid *m/z* of each peptide. In HDX-Analyzer, multiple linear regression is used to derive a model of deuterium uptake over time for a peptide of an experimental state. A paired Student's t-test or analysis of covariance (ANCOVA) is used to determine whether a statistically significant difference exists between the two states. In the ANCOVA model, experimental states are treated as groups, with the deuterium uptake or centroid *m/z* as the dependent variable, and time as a covariate that correlates with the dependent variable but is not a focus of the study. In ANCOVA, the potential effects of the covariate are used to adjust the mean value of the dependent variable, such that the effects of the covariate are minimised. In both paired Student's t-test and ANCOVA, the point estimate of the mean difference in uptake between two peptides is calculated, along with its confidence intervals and p-value for a given confidence level. These values are presented for each peptide and allow the user to assess statistical significance from multiple angles. Pairwise t-tests however suffer from the multiple comparisons problem in which the intended confidence level is compounded with each subsequent comparison. For example, three pairwise comparisons each with 0.95 confidence, results in an actual confidence level of 0.857 ($0.95 \times 0.95 \times 0.95$) and a final confidence level of 85.7%. The unintended shift in the confidence level leads to potential type I errors or false positives in which a true null hypothesis (no change) is rejected. To avoid the multiple comparisons problem with

pairwise t-tests, multiple comparisons can instead be performed using other statistical methods such as the Tukey's range test as seen in HDX Workbench[122] or employ the Bonferroni[k] correction method[147].

Instead of using multiple linear regression for modelling of deuterium uptake over time, MEMHDX first fits the deuterium uptake data using a linear mixed effects model. The mixed effects model is interesting as it treats measurement variance over replicates, as a random effect on the deuterium uptake. In the mixed effects model, replicates can be seen as random effects on the true mean of the measurement, as in theory, there can be an infinite number of replicate measurements, each taken at different conditions leading to "random" variations across the measurement that need to be accounted for in the statistical model. The utilisation of a mixed effects model to estimate the deuterium uptake over time is a significant upgrade to the DynamX methodology in which uptake is taken as the simple intensity weighted average of the replicates composing the peptide. Deuteros 2.0 also follows DynamX and no statistical models are employed to attempt at accounting for differences across the measurement replicates. To determine whether a statistically significant difference exists between two peptides, MEMHDX calculates two P-values using a chi-squared test. Similar to HDX-Analyzer, the P-values of MEMHDX attempt to take into account the effect of time in the deuterium uptake measurement. This differs the P-value calculated in the Deuteros 2.0 volcano plot, which does not take into account the acquisition time. In MEMHDX, the $\text{P-value}_{\text{Magnitude\_of\_Delta\_D}}$ measures the statistical significance of deuterium uptake difference between the experimental conditions. The $\text{P-value}_{\text{Change\_in\_Dynamics}}$ tests whether there is an observed change in deuterium uptake over time (hence uptake dynamics) between the two states and thus accounts for the time axis.

k The Bonferroni method corrects for type 1 errors (false positives) in multiple comparisons by adjusting α of individual statistical tests, such that their combined significance is no more than the intended significance ($\alpha' = \alpha/n$).

3.5.5 Statistical filtering

Statistics are applied to differential HDX-MS data ultimately with the aim of determining whether the observed deuterium uptake between two experimental states, is sufficiently large to yield statistical significance. The elaborate methods of applying statistics essentially allow a binary decision on whether or not a peptide has behaved differently in a different condition. In HDX-Analyzer, the user is left with a statistical breakdown of the comparison between states, and no conclusions are made for the user. A simpler method of making this decision compared to those of MEMHDX and HDX-Analyzer, is the confidence interval first conceived by Houde *et al*. which has been used in Deuteros 2.0. Specifically, this confidence interval is used to represent the expected range of random variation for replicate measurements and is calculated using the replicate variance. In practice, it is applied as a range around $y = 0$ representing the sample noise of the measurement. Thus, deuterium uptake differences outside of this range are designated as statistically significant. Also different, is that this confidence interval is measured across all peptides and is globally applied to each timepoint individually, rather than a per-peptide basis such as in ANCOVA or t-test methods. This time-centric application has led to Deuteros 2.0 to focus on a per-timepoint representation of the differential HDX-MS data, rather than per-peptide or per-state. The confidence interval of Deuteros 2.0 will be referred to as the "global confidence interval". Using the global confidence interval approach, peptides are grouped into one of three categories: deprotected, protected and those with non-significant changes. The only other software to perform peptide grouping is MEMHDX.

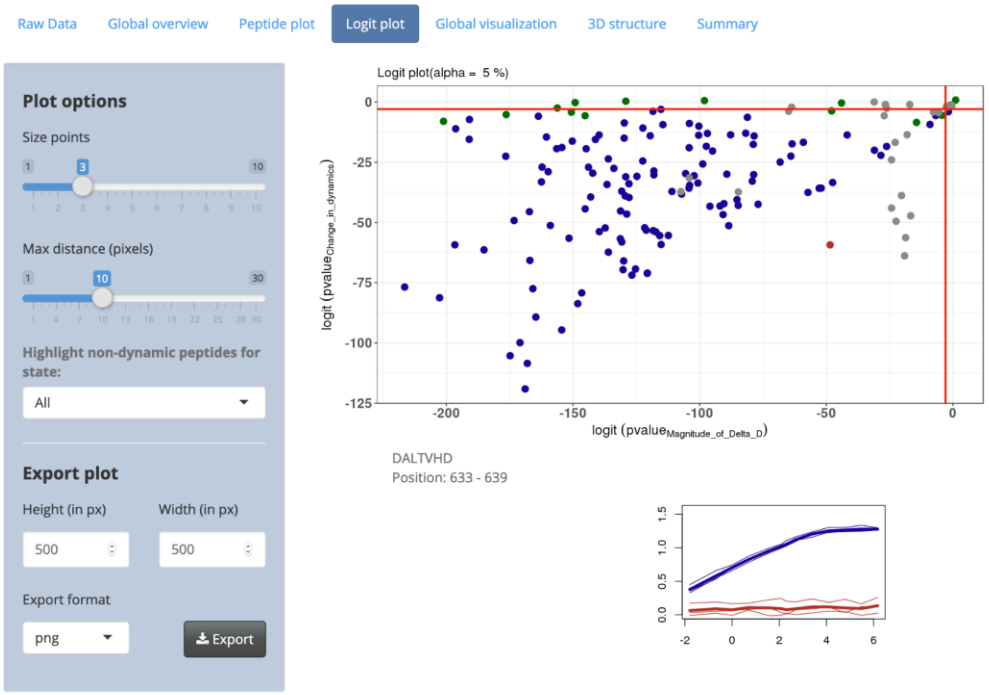


Figure 3.24. The MEMHDX Logit plot feature. Left shows the options panel for controlling the Logit plot. Right shows the logit plot. Only customisations of the plot size and datapoint sizes are available in MEMHDX. Peptides can be highlighted in the Logit plot according to their classification in one of four categories. Red and blue peptides are those that are deprotected and protected in the holo state compared to the apo state. Grey and grey peptides show high or no dynamics in both states. Hovering over the single red data point at approximately (-50,-60) spawns the uptake plot at the lower right of the window.

In MEMHDX, plotting the *logit* function of each P-value, *logit*(P-value$_{Magnitude_of_Delta_D}$) and *logit*(P-value$_{Change_in_Dynamics}$) allows clustering of peptides into one of four categories (Figure 3.24): 1) peptides showing increased uptake in the holo state compared to apo (equivalent to deprotected), 2) peptides showing decreased uptake in the holo state compared to apo (equivalent to protected), 3) peptides with no differences between apo and holo (non-significant) and 4) peptides dynamic in both apo and holo. The existence of peptides of the fourth category are not identified in Deuteros 2.0 and inclusion of this feature should be considered in future updates. Categorisation of peptide behaviour is through plotting the *logit* function of P-values, *logit*(P-value$_{Magnitude_of_Delta_D}$) against *logit*(P-value$_{Change_in_Dynamics}$) in

MEMHDX. A combination of the mixed effects model, dual P-values and category separation makes MEMHDX the most statistically stringent HDX-MS software currently available.

Statistical analysis of peptide ensembles is an important feature of the HDX-MS technique and over the recent years, many systems have been elucidated using these methods. The global confidence interval approach introduced by Houde *et al*. (2011) is a popular method of statistical filtering of peptide data and have been used for the characterisation of systems including antibody flexibility[148], antibody-drug conjugates[149], lipid dependent effects on nanodisk scaffolds[150] and secondary transporters[141], bacterial translocases[151] and lanthipeptide synthetases[152]. In these studies, authors either calculate the corresponding value of the global confidence interval for their system of acquisition using the equation (3.13), or directly apply the ± 0.5 limit detailed in the original Houde *et al*. publication[140]. Interestingly, the study by Habibi *et al*. utilised Deuteros only for calculation of the 99% confidence interval but did not use any of the plots generated by the software[152]. Instead, the confidence interval value was taken from the software and transplanted onto the difference plot produced by DynamX, potentially illustrating that some users may prefer a representation of peptide data that is not separated by timepoints. This should be kept in mind for future implementations of Deuteros 2.0.

Recent studies that have utilised the statistical analysis features of MEMHDX include those that characterise protein-protein interactions involving monoclonal antibodies[153], the CRISPR Cascade complex[154], PPARγ/RXRα nuclear receptor transcription factors[155] and the bacterial membrane secretin PulD[156]. In each of these publications, Logit plots and the corresponding peptide categories were not used for biological interpretations. Similar to that observed by Habibi *et al*., Terral *et al*. displays the differential HDX-MS data in the difference plot format and interestingly include the global confidence interval for significance testing - however this is function is not performed by MEMHDX and the authors do not elaborate the method

of their analysis. As such, while we may discuss the relative usefulness of either Deuteros 2.0 or MEMHDX, it is clear from this small sample of publications that the focus of the software features should be around users and their analysis needs, and not which software provides more and advanced functions.

Chapter 4: Dynamic characterisation of large protein complexes: the COP9 Signalosome

Preface

The following publication has been presented in this chapter in the '*Thesis Incorporating Publications*' format:

Faull, S. V.[†], Lau, A. M. C.[†], Martens, C., Ahdash, Z., Hansen, K., Yebenes, H., Schmidt, C., Beuron, F., Cronin, N. B., Morris, E. P., Politis, A. (2019). Structural basis of Cullin 2 RING E3 ligase regulation by the COP9 signalosome. *Nature Communications,* 10, 3814, doi:10.1038/s41467-019-11772-y.

[†] Denotes authors of equal contribution.

Author contributions

As co-first author of Faull and Lau *et al*. 2019[157], I performed all modelling for this project, including fitting structures of the CSN and CRL2 complexes into cryo-EM maps. I performed all analysis and interpretation of MS data (including native MS, XL-MS and HDX-MS). I also performed all sample preparation (downstream of protein purification) of CSN and CRL2 samples for acquisition of each set of MS data. I led the publication writing process and produced all figures for the article.

All samples of CSN and CRL2, and cryo-EM maps used for this study were kindly provided by Dr Sarah Faull who was assisted by Dr Hugo Yebenes, Dr Fabienne Beuron, Dr Nora Cronin and Dr Ed Morris. Dr Chloé Martens performed all non-PLIMSTEX HDX-MS data acquisition and is responsible for the native spectra of the

CSN-CRL2 complexes. The native spectra of apo-CSN was collected by Dr Argyris Politis. Zainab Ahdash is responsible for collection of all IM-MS data. Kjetil Hansen collected all PLIMSTEX data. XL-MS data were acquired by Dr Carla Schmidt.

4.1 Extended introduction into the biology of the COP9 Signalosome and its role as the regulator of NEDD8-activated CRL E3 Ligases

4.1.1 Ubiquitin-Proteasome System

The ubiquitin-proteasome system (UPS) comprises an impressive multicomponent network that regulates an extensive number of critical biochemical pathways ranging from cell-level protein turnover, cell signalling, DNA repair to organism-level processes such as angiogenesis and embryonic development[158] (Figure 4.1). The discovery and mapping of the UPS network and its role in mediating cellular protein degradation, warranted the 2004 Nobel Prize in Chemistry award to Aaron Ciechanover, Irwin Rose and Avram Hershko.

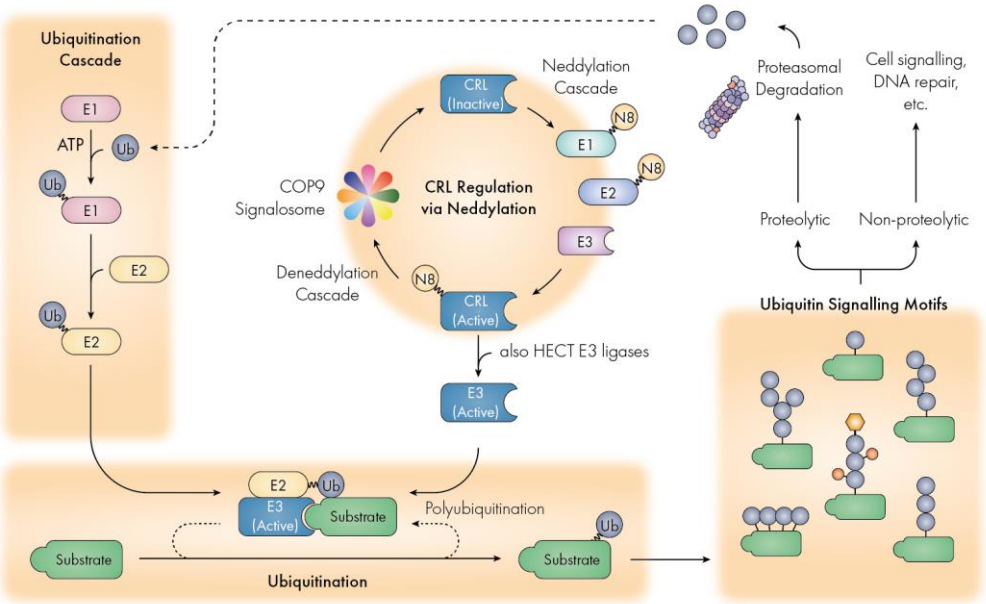


Figure 4.1. A simplified view of the Ubiquitin-Proteasome System. This chapter will discuss the different branches of the UPS including ubiquitination leading to proteasomal degradation and regulation of key effectors of the system known as Cullin RING ligases (CRLs).

The turnover of proteins within cells is modulated through a highly regulated partnership involving two key players: the proteasome - a large 2 MDa complex that contains a myriad of enzymatic centres for endoproteolysis, and ubiquitin (Ub) - a

conversely small 8 kDa globular protein that functions as protein tag. Cellular proteins can undergo one of many post translational modifications, one of which is ubiquitination which involves the covalent ligation of ubiquitin via its C-terminal G76, onto a lysine residue of the substrate protein[159]. Ubiquitination of substrate proteins chaperone them into a number of downstream cellular fates, including protein degradation by the proteasome, sub-cellular trafficking and other signalling pathways. The process of substrate ubiquitination is catalysed by a well-defined cascade of three classes of proteins, termed the E1 ubiquitin activating enzymes, E2 ubiquitin conjugation enzymes and E3 ubiquitin ligases. The E1, E2 and E3 cascade is established in a pyramidal scheme whereby in humans, only one to two E1 enzymes are responsible for mobilisation of up to 40-100 E2 enzymes, which are in turn utilised by approximately 10^3 E3 ligases[160-162]. Similar diversity of E1-E3 enzymes have also been identified across different plant and yeast species[163].

In the first step of the ubiquitination cascade, a molecule of ubiquitin is fused to an E1 enzyme in an ATP-dependent reaction (Figure 4.2a). The E1 with its covalently conjugated ubiquitin, E1~Ub (where the "~" stylization denotes a covalent bond), next transfers ubiquitin to an E2 enzyme, forming an E2~Ub complex (Figure 4.2b). Most of the cellular E2 pool exists in the E2~Ub format, ready for utilisation by E3 enzymes[164]. In the final step, an E3 ubiquitin ligase docks with the substrate protein and the primed E2~Ub forming the ubiquitination machinery (Figure 4.2c).

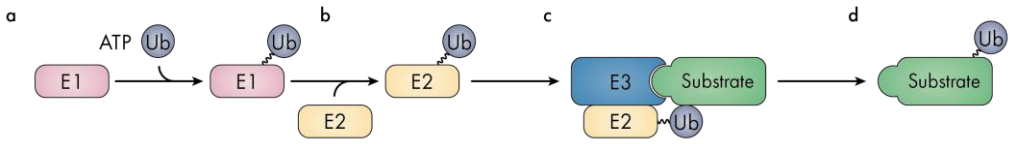


Figure 4.2. Ubiquitination cascade. (a) An E1 ubiquitin activating enzyme in an ATP-dependent step, covalently bonds with a molecule of ubiquitin. (b) Ubiquitin from the E1~Ub complex is transferred to an E2 ubiquitin conjugation enzyme, forming E2~Ub. (c) E2~Ub and substrates dock with an E3 ubiquitin ligase which catalyses the transfer of ubiquitin from the E2 enzyme to the substrate. (d) Ubiquitination results in biochemical changes in the substrate that leads to a myriad of cellular fates. Figure adapted from Rape, 2018 [159].

4.1.2 The Ubiquitin Code

Besides ubiquitination of substrate molecules, ubiquitin itself can also be ubiquitinated at one of eight different residues - at seven lysines (K6, K11, K27, K29, K33, K48 and K63) and its N-terminal M1 methionine. Subsequent polyubiquitination can also occur, leading to elongation of the chain to more than 10 units of ubiquitin long[165]. Polyubiquitination at specific lysine linkages (e.g. K48 or K63) chaperone the substrate protein to specific cellular fates[159]. The combination of various ubiquitin chain lengths, topologies and modifications, impressively diversifies the possible array of ubiquitination motifs, leading to a highly complex and specific signalling system. In fact, entire sets of enzymes have been discovered that are capable of reading and editing this "ubiquitin code"[165].

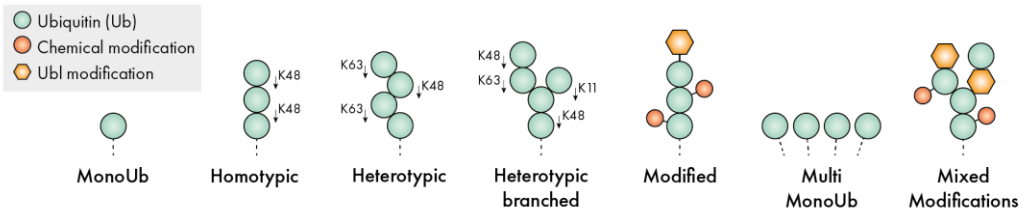


Figure 4.3. Ubiquitin signalling motifs. Several different flavours of ubiquitination motifs are possible, including monoubiquitination and polyubiquitination. Successive polyubiquitination can involve the same lysine linkage (homotypic) or different lysine linkage (heterotypic) which can also be branched. Ubiquitin chains can be modified through different chemical modifications such as phosphorylation or acetylation, or ligation of other ubiquitin-like (Ubl) modifiers such as SUMO or NEDD8.

Multiple rounds of successive ubiquitination can lead to chains of various lengths and topologies (Figure 4.3). Figure 4.4 demonstrates the ubiquity of the ubiquitin signalling network.

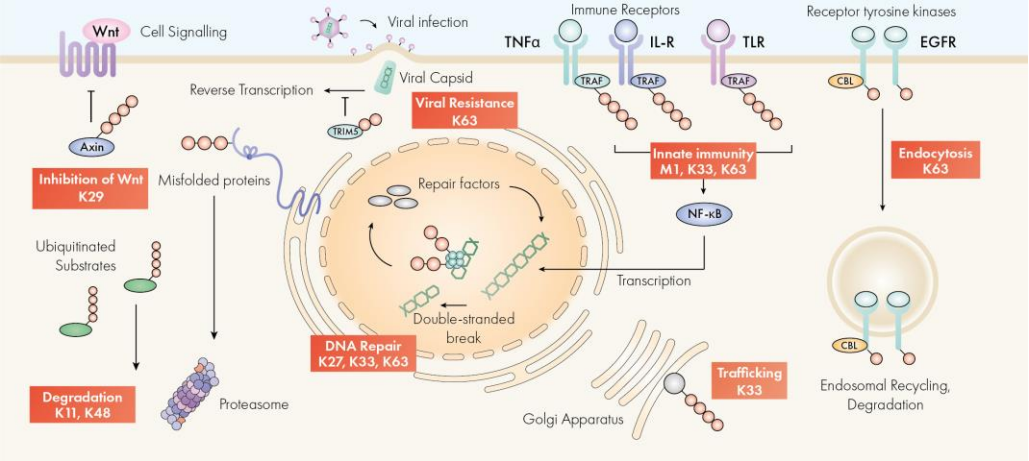


Figure 4.4. Cellular roles of ubiquitin signalling. Various signalling processes and pathways are regulated via ubiquitin signalling, including ubiquitin-dependent proteolysis of expired or misfolded proteins, trafficking of post-Golgi or cell surface receptors, DNA repair and inhibition of cell signalling pathways such as Wnt.

4.1.3 The Proteasome

Proteolytic signals such as polyubiquitin linkages via K48 are recognised by the cell's recycling powerhouse, known as the proteasome (Figure 4.4). Two flavours of proteasomes are found within the cell, the 26S and 30S, named so due to their sedimentation coefficients measured in Svedbergs[166]. The 30S is the functional variant of the proteasome, capable of proteolysis, while the 26S is in fact a partial subcomplex of the 30S. Readers may however find in literature that the naming of the 26 and 30S proteasome is controversial. The reason for this is historic[167]. Attempts at measuring the sedimentation coefficient of the fully assembled 30S using density-gradient centrifugation were in made on the 26S partial complex[168], leading to the misnomer that the 26S is functionally active. Later biophysical analysis confirmed that the active proteasome has a sedimentation coefficient of 30S and the 26S is likely a partial complex[166,167].

The 30S proteasome is formed by a 20S catalytic core particle and capped at both ends by 19S regulatory subunits (Figure 4.5). The core particle is a large barrel-like structure consisting of four stacks of α and β heptameric rings in an αββα configuration. Within the internal cavity, the 20S core contains several enzymatic

sites designed to hydrolyse peptide bonds of acidic, basic and hydrophobic residues.

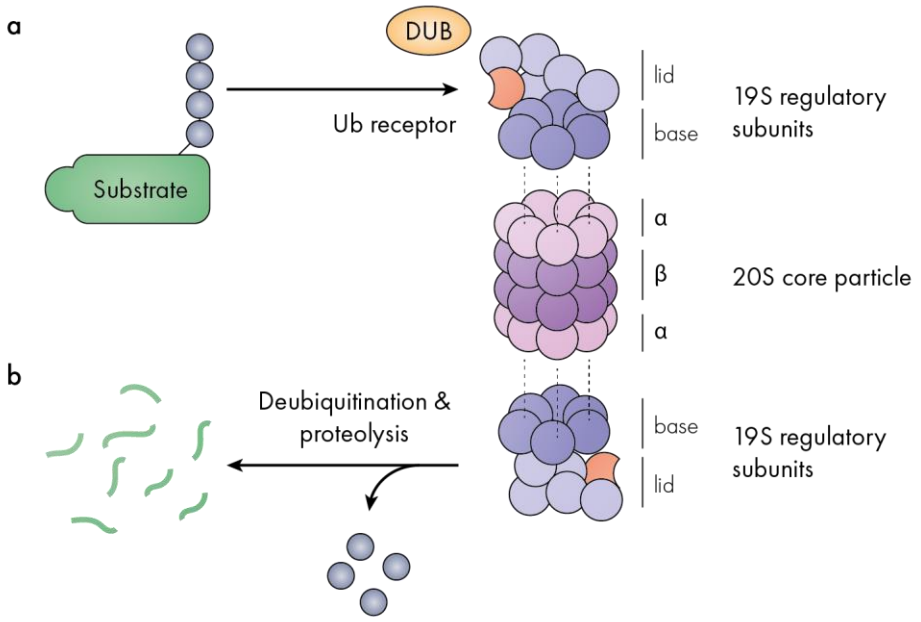


Figure 4.5. Protein degradation by the 30S proteasome. The 30S proteasome is composed of a 20S core and two 19S regulatory particles. (a) Tetraubiquitinated substrates can be recruited to the 19S ubiquitin receptor, along with deubiquitinases. (b) Ubiquitin is cleaved from the substrate via proteolytic activity of the deubiquitinase, and the substrate is digested by the 20S core particle.

In its isolated αββα configuration, the 20S is auto-inhibited due to the enzymatic sites being located at the β core of its cavity of which access is sterically restricted by the α7 rings. To activate the 20S, two 19S regulatory complexes are recruited to either end of the 20S (Figure 4.5). The 19S particles consist of a base and lid subcomplexes which contribute a variety of biochemical functions to the proteasome including receptors specific to substrates with particular ubiquitination motifs. In a series of steps, ubiquitin are stripped off the substrate by recruited deubiquitinases, the substrate is unfolded and lysed into oligopeptides by the catalytic centres of the 20S core. Free ubiquitin can then be recycled back into a number of biochemical pathways (Figure 4.4).

4.1.4 E3 Ubiquitin Ligases

E3 ligases provide the ubiquitination cascade with substrate specificity and are involved in a number of highly important biochemical pathways including progression of the cell cycle, immune response and protein degradation via ubiquitination[169]. There are three currently identified classes of E3 ligases known as RING, HECT and RBRs.

By far the most prominent family of E3 ligases with more than 600 members in mammals, are those that contain the RING or Really Interesting New Gene domain. The RING domain is a highly conserved and characteristic zinc-binding fold that is capable of associating with or allosterically activating E2~Ub complexes. RING E3 ligases may function as monomeric, homo- and heterodimeric or multi-subunit complexes. Monomeric RING E3s include CBL that ubiquitinates activated receptor tyrosine kinases, triggering their endocytosis, or the TRAF family of E3 ligases that autoubiquitinate following activation of immune receptors (Figure 4.4) and lead to activation of the NF-κB pathway. Multi-subunit RING E3s include the Cullin RING E3 ligases (CRLs) and also two related but otherwise distinct complexes - the APC2 subunit of the anaphase promoting complex/cyclosome (APC/C) and the p53 anchoring protein complex (PARC, also known as CUL9).

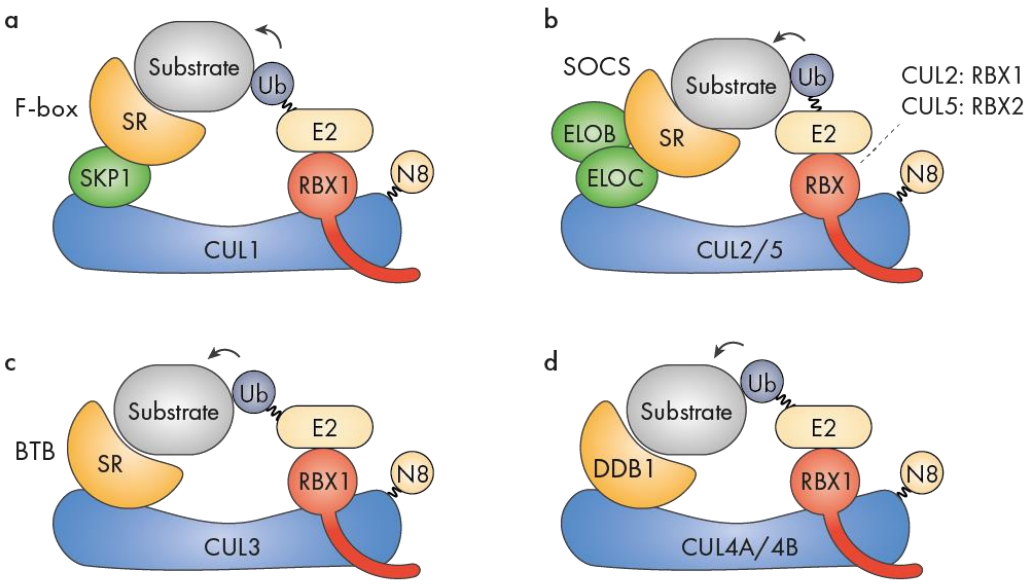


Figure 4.6. Organisation of the modular Cullin-RING E3 ligases. Schematics depict the subunit layout for (a) CRL1 and its F-box binding substrate receptors, (b) SOCS-box CRL2 and CRL5, (c) BTB binding CRL3 and (d) CRL4A/4B in complex with its DDB1 substrate receptors. Cullins are shown in blue, RBX E3 in red, adaptors in green and substrate receptors in orange. CRLs must first be neddylated involving covalent ligation of NEDD8 (N8) before the E3 ligase activity of RBX subunits are active.

CRLs are the largest family of E3 ligases consisting of modular scaffolds with E3 ligase activity[170]. It is thought that CRLs may mediate up to 1/5th of the cellular protein turnover via the proteasome[171]. Though there are exceptions, each CRL typically consists of several components: an elongated and curved Cullin (CUL) scaffold protein, a RING-box protein (RBX) that functions as the E3 ligase of the complex, and adaptor proteins and substrate receptors which provide substrate specificity (Figure 4.6). So far, seven different Cullin scaffolds that have been identified - CUL1, 2, 3, 4A, 4B, 5 and 7. All Cullins with the exception of CUL5 associate with RBX1, while a homologous RBX2 binds CUL5 exclusively (Figure 4.6). At the minimum, a Cullin must be in complex with its RBX E3 ligase for it to be termed a CRL, for example, CUL1-RBX1 is termed CRL1, CUL2-RBX1 is CRL2 and so on. Each CRL is capable of associating with an individualised array of adaptor proteins and substrate receptors, totalling more than 400 interchangeable components[172]. This modularity provides

the UPS network with tremendous specificity for over 1000 substrates in a highly regulated system[169,173].

Table 4.1. Crystallographic structures of CRL complexes.

System	PDB	Resolution (Å)	Subunits	Year	Reference
CRL1	1LDJ	3.00	CUL1, RBX1	2002	Zheng *et al*.[174]
	1LDK	3.20	CUL1, RBX1, Skp1, Skp2	2002	Zheng *et al*.[174]
	1U6G	3.20	CUL1, RBX1, TIP120	2004	Goldenberg *et al*.[175]
	3RTR	3.21	CUL1, RBX1	2011	Calabrese *et al*.[176]
	3TDU	1.50	CUL1, UBC12 (E2), DCNL1	2011	Scott *et al*.[177]
	3TDZ	2.00	CUL1, DCNL1, Peptide	2011	Scott *et al*.[177]
	4F52	3.00	CUL1, RBX1, Glomulin	2012	Duda *et al*.[178]
	4P5O	3.11	CUL1, RBX1, DCNL1, UBC12 (E2), NEDD8	2014	Scott *et al*.[179]
	5V89	1.55	CUL1, DCNL4	2017	Scott *et al*.[180]
CRL2	4WQO	3.20	CUL2, ELOB, ELOC, VHL	2015	Nguyen *et al*.[181]
	5N4W	3.90	CUL2, RBX1, ELOB, ELOC, VHL	2017	Cardote *et al*.[182]
CRL3	4EOZ	2.40	CUL3, SPOP	2012	Errington *et al*.[183]
	4AP2	2.80	CUL3, KLHL11	2013	Canning *et al*.[184]
	4APF	3.10	CUL3, KLHL11	2013	Canning *et al*.[184]
	6I2M	2.30	CUL3, A55	2018	Gao *et al*.[185]
	4HXI	3.51	CUL3, KLHL3	-	To be published
	5NLB	3.45	CUL3, KEAP1	-	To be published
CRL4A	2HYE	3.10	CUL4A, RBX1, Viral protein	2006	Angers *et al*.[186]

	4A0K	5.93	CUL4A, RBX1, DDB1, DDB2	2011	Fischer *et al*.[187]
CRL4B	4A0C	3.80	CUL4B, RBX1, CAND1	2011	Fischer *et al*.[187]
	4A0L	7.40	CUL4B, RBX1, CAND1	2011	Fischer *et al*.[187]
	4A64	2.57	CUL4B	-	To be published
CRL5	3DPL	2.60	CUL5, RBX1	2008	Duda *et al*.[188]
	3DQV	3.00	CUL5, RBX1, NEDD8	2008	Duda *et al*.[188]
	4JGH	3.00	CUL5, ELOB, ELOC, SOCS2	2013	Kim *et al*.[189]
	4N9F	3.30	CUL5, ELOB, ELOC, CBFB, Vif	2014	Guo *et al*.[190]
CRL7	-	-	-	-	-

Most of the current information known about CRLs stem from the early discovery of the archetypal CRL1[191,192]. Complete and partial crystal structures of each of the CRLs soon followed over the next two decades, including complete structures of the CRL1, CRL2, CRL3, CRL4A, CRL4B and CRL5. A list of CRL crystal structures can be found in Table 4.1. Among the 26 crystal structures of CRLs solved thus far, only CRL7 has no structural representation. The intact structures of CRL3 and CRL5 have also not yet been determined, but only N or C-terminal fragments. Comparisons of the CRL1-5 structures reveal a highly conserved Cullin topology consisting of seven domains (Figure 4.7). The N-terminal stalk is formed from three five-helix bundles known as Cullin-Repeat domains (CR1-3). A four-helix bundle connects the CR1-3 domains to the Cullin C-terminal which is formed from the Cullin homology C-terminal domain (CTD) and the Winged-Helix A and B domains (WHA/WHB) (Figure 4.6). CRLs can be divided into two portions - the N-terminal and C-terminal regions. The N-terminal CR1 domain functions as a substrate receptor complex binding site, providing specificity to substrates and allowing them to dock onto the N-termini of the CRL. At

the opposing C-terminal domain, the RBX subunit which functions as an E3 ligase, awaits for specific regulatory signals before recruiting a E2~Ub complex. In a lesser understood final step, ubiquitin is transferred from the E2 enzyme to the substrate and this marks the completion of one round of ubiquitination.

Table 4.2. Adaptors and receptors for CRL complexes.

System	RING	Adaptors	Receptor motif
CUL1	RBX1	Skp1	F-box
CUL2	RBX1	ELOB/ELOC	BC-box
CUL3	RBX1	-	BTB
CUL4A	RBX1	-	DCAF
CUL4B	RBX1	-	DCAF
CUL5	RBX2	ELOB/ELOC	SOCS-box
CUL7	RBX1	Skp1	F-box

Which adaptors and substrate receptors are capable of binding to the N-terminal of CRLs, depends on the flavour of Cullin. Several common binding motifs have been identified between the Cullin family members. These are summarised in Figure 4.6 and Table 4.2. In the archetypal CRL1, the substrate adaptor is the S-phase kinase-associated protein 1 (Skp1) which contains an F-box binding site[170]. The binding site provides CRL1 with modular access to F-box domain-containing substrate receptors such as Skp2 and Fbw7[193], in turn connecting CRL1 with its important cognate substrates such as p53 (via Skp2)[194] and cyclin E (via Fbw7)[195]. Both CRL2 and CRL5 utilise the identical substrate adaptor proteins Transcription Elongation Factor B Polypeptides 1 and 2, also known as Elongin C (ELOC) and Elongin B (ELOB) respectively. Surprisingly, while ELOB and ELOC bind to both CRL2 and CRL5, their presence does not provide the Cullins with ubiquitous substrate specificity[196]. In fact, crystal structures of the CRL5 in complex the viral infectivity factor (Vif) receptor (PDB 4N9F), revealed that a three-way interaction network occurs between the interfaces of CUL5, ELOB/C and Vif. Taken together these suggest that substrate specificity may

be determined at both the level of the substrate adaptor and the Cullin family member[169]. Receptors that bind to CRL2 are ELOB/ELOC (BC)-box binding proteins while those associated with CRL5 contain the suppressors of cytokine signalling (SOCS)-box[169]. Unlike the earlier CRLs, CRL3, CRL4A and CRL4B do not require substrate adaptor molecules. Instead, their receptors directly associate with their N-termini. In CRL3, the binding motif is known as the Broad complex, Tramtrack, Bric-a-brac (BTB) domain. CRL4A and 4B both associate with the DNA damage binding protein 1 (DDB1) coupling them to the DNA repair response and for some time their receptor motif was unknown[169,170]. Further studies identified over 50 candidate DDB1-binding proteins with a shared binding motif that is now known as the DDB1-CUL4-associated factors (DCAF)[197].

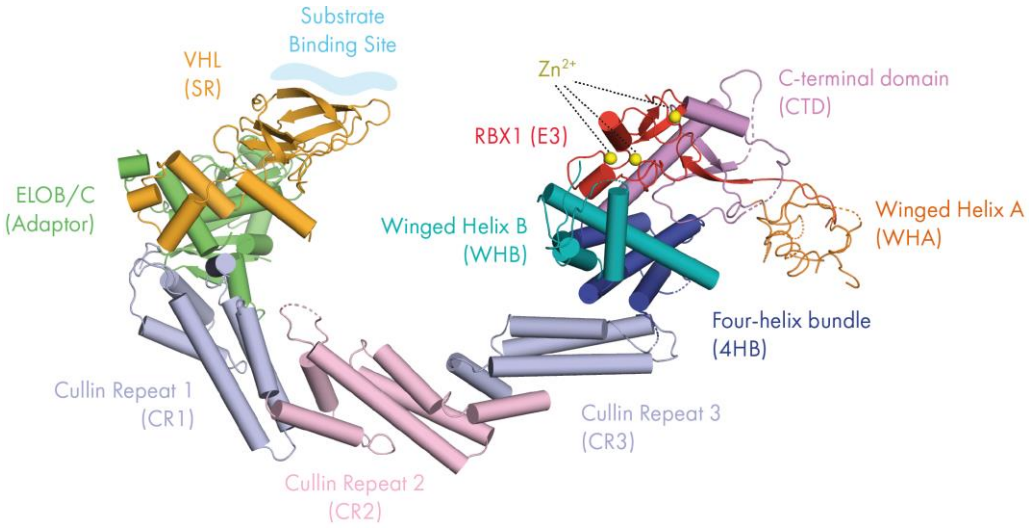


Figure 4.7. Domain and subunit structure of the CRL2 E3 ligase. The structural of CRL2 can be used to generalise the topology of all CRLs. CUL2 contains seven domains: CR1-3, 4HB, CTD (Cullin homology domain) and WHA/B. The substrate adaptors ELOB and ELOC (green) allow the substrate receptor (SR) VHL (orange) to associate with the N-terminal CR1 domain of CRL2 (light purple). The RBX1 E3 ligase is cradled by the WHB domain. Model generated from a combination of both PDBs 5N4W (CUL2, RBX1) and 4WQO (VHL, ELOB and ELOC).

4.1.5 Biology of the Cullin 2-RING E3 Ligase

As one of the most biochemically but least structurally characterised CRLs, the CRL2 regulates a number of important cellular growth factors, including angiogenesis, viral resistance and cell migration[196]. The CRL2, through its substrate adaptors ELOB and ELOC, interacts with at least five BC-box binding receptors[198] including VHL[a], LRR-1[b], FEM1[c], PRAME[d] and ZYG11[e]. All receptors interacting with the BC-box of ELOB and ELOC, further contain a Cullin-2 box that confers specificity to the CUL2 scaffold. Collectively, the BC and Cullin-2 boxes are known as the VHL box and is found in all CUL2-binding receptors[198]. Unsurprisingly, the VHL box was first identified in the VHL substrate receptor. In complex with VHL, the $CRL2^{VBC}$ (where VBC represents VHL-ELOB-ELOC) regulates modulates the angiogenesis pathway via regulation of hypoxia inducible factor (HIF)[196,198]. The angiogenesis pathway oversees the regulation of biochemical processes that leads to the formation of new blood vessels[199]. HIF is a heterodimeric transcription factor composed of unstable HIF1-α and constitutively expressed HIF1-β. In hypoxic conditions (low oxygen), HIF1-α binds to HIF1-β, forming active HIF which then translocates to the nucleus where it binds to and activates DNA promoters known as hypoxia response elements (HRE). HREs are upstream of angiogenic genes including vascular endothelial growth factor (VEGF), glucose transporter 1 (GLUT1) and platelet-derived growth factor (PDGF)[196]. In normoxic conditions (normal oxygen), HIF1-α is hydroxylated at specific prolines by a class of prolyl hydroxylases, leading to recognition by VHL, ubiquitination and proteasomal degradation[196,198]. Mutations in VHL, cause the rare hereditary von Hippel-Lindau cancer syndrome which leads to uncontrolled vascular tumours due to non-regulation of angiogenesis and other cell growth pathways[196].

---

a VHL: von Hippel Lindau

b LRR-1: Lysine-rich repeat 1

c FEM1: Feminization 1

d PRAME: Preferentially expressed antigen of melanoma

e ZYG11: Acronym not defined in literature

Studies have also explored the possibility of taking advantage of the high specificity of CRLs, coupling them target and degrade certain proteins contributing to pathology or cellular malfunction. These are so called proteolysis targeting chimeras (PROTAC), which are synthetic chimeric molecules that show specificity for an E3 ligase substrate receptor, and a target protein (Figure 4.8)[193]. The first PROTAC was developed by Sakamoto *et al*. who showed that the estrogen receptor could be preferentially degraded *in vitro* by designing a PROTAC that could connect between the receptor and the CRL1 E3 ligase[200,201]. Later in 2004, the first successful *in vivo* use was demonstrated when the $CRL2^{VBC}$ was coupled to the estrogen receptor[202]. Since then, an overwhelming number of PROTAC successes have been developed using the $CRL2^{VBC}$ system[201]. Despite intense research interest and numerous structural models of CRLs provided through x-ray crystallography, the underlying conformational dynamics which allow CRL assemblies, including the CRL2, to perform their biological function, remain ambiguous. Given that there are differences between members of the CRL family, inferences may not be directly made between CRL2 and other CRLs. The emerging field of PROTACs which interact through VHL, also justifies the need for proper characterisation of the CRL2 system.

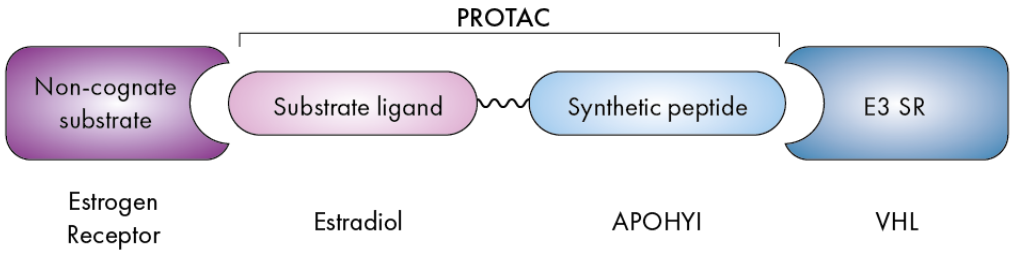


Figure 4.8. PROTAC layout. PROTACs couple E3 ligases with non-cognate substrates using a synthetic molecule consisting of moieties that bind to both the E3 substrate receptor and the intended substrate. Coupling results in E3-dependent ubiquitination and proteasomal degradation.

4.1.6 Regulation of Cullin RING E3 Ligases

The regulation of CRLs involves three different but related mechanisms: 1) neddylation, 2) deneddylation and 3) sequestration (Figure 4.9)[203,204]. Neddylation is a post-translational modification that like ubiquitin, consists of a three-tiered enzymatic cascade of E1, E2 and E3 proteins. In neddylation, a molecule of the 8 kDa UBL NEDD8, is covalently conjugated via its C-terminal G76 to a conserved acceptor lysine on Cullin proteins located within the WHB domain. Neddylation results in an isopeptide bond formed between the sidechain of the conserved lysine Cullin and NEDD8. Key stages of the neddylation reaction have been captured in CRL1 and CRL5 systems via crystallography, allowing imagination of how neddylation occurs in order to activate CRLs. The current and widely recognised neddylation mechanism is displayed in Figure 4.9a.

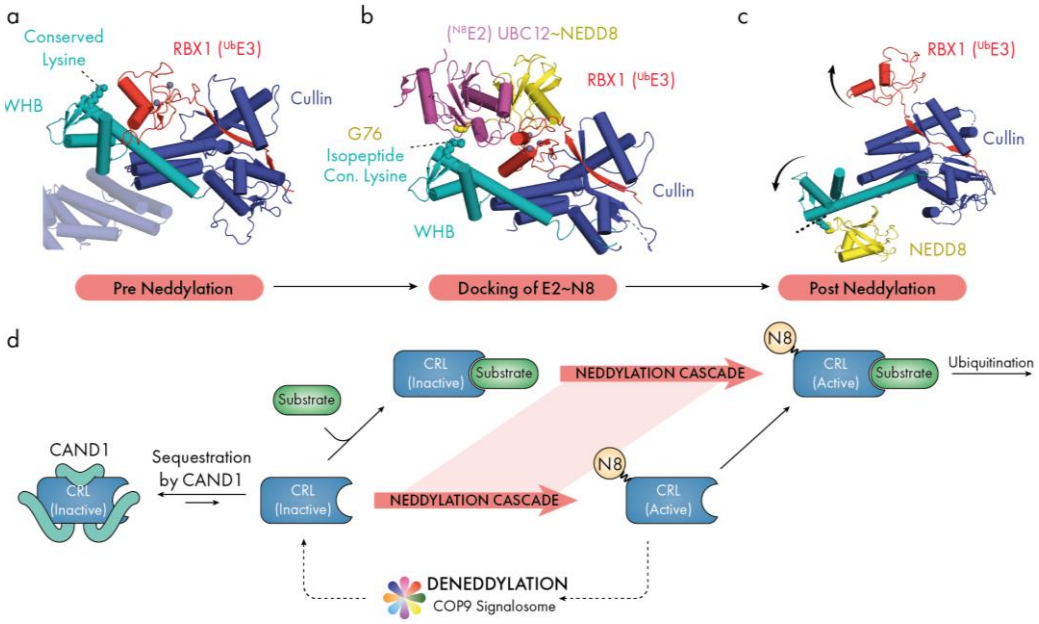


Figure 4.9. Key regulators of CRL activity. Various stages of the neddylation reaction have been accessed via X-ray crystallographic structures. (a) Pre-neddylated CRL1 (PDB 1LDJ) features the ubiquitin E3 ligase (UbE3) RBX1 inhibited against the WHB of CUL1. (b) The neddylation machinery consists of the N8E2 enzyme UBC12 and conjugated NEDD8 and the DCNL1 E3 ligase (not shown) which ligates NEDD8 via G76 onto the conserved lysine of CUL1 and forming an isopeptide bond (PDB 4P5O). (c) Dissociation of NEDD8 E2 and E3 enzymes results in the post neddylation conformation of CRL5 (3DQV) in which WHB~N8 is found dramatically rotated away from its inactive position in (a). RBX1 is conformationally less restricted following neddylation. (d)

Various modes of CRL regulation. Cellular CRLs are thought to be sequestered by CAND1. Binding of substrates or dissociation of CAND1 may lead to neddylation of the CRL2. Neddylation is reversed by the deneddylase activity of the COP9 Signalosome.

In a non-neddylated state, the E3 ubiquitin ligase RBX1 is cradled against the WHB domain of Cullin and is inactive (Figure 4.9a). The role of the WHB in inhibiting E3 activity of RBX is supported through mutagenesis studies which observed that a single residue substitution (Y761A) in the WHB domain of CUL1 was sufficient to promote neddylation-independent E3 activity[205]. The cellular pool of inactive non-neddylated CRLs are also competed for by the 120 kDa Cullin-Associated NEDD8-Dissociated 1 (CAND1) protein. Crystal structures of CAND1 in complex with CRL1 (1U6G) and CRL4B (4A0C), revealed an elongated brace structure that clamps onto Cullin along its full length and sterically obstructs both N-terminal substrate and C-terminal E2 binding sites[175]. CAND1 acts as a positive regulator of CRLs since sequestration prevents both auto-ubiquitination of substrate receptor complexes and also enhances their substrate receptor exchange of CRLs[206].

Only when not sequestered by CAND1, can CRLs undergo neddylation. Drawing many parallels with ubiquitination, neddylation first involves the E1 NEDD8-activating enzyme, also referred to as NAE[207]. NAE auto-conjugates via one of its reactive cysteine side chains with NEDD8 in an ATP-dependent step, forming a high energy thioester bond. NEDD8 is then transferred from NAE~NEDD8 to the E2 UBC12, which together with the E3 Defective in Cullin Neddylation 1 (DCN1), transfers NEDD8 to the acceptor lysine of Cullins (Figure 4.9b). Although neddylation is promoted in the presence of DCN1, its presence is not strictly necessary in vivo, leading to suggestions that RBX can function as a NEDD8 E3 ligase while DCN1 is a co-factor for neddylation[208]. Crystal structures of the neddylation intermediate complex for CRL1 (PDB 4P5O), revealed that binding of UBC12~N8 to CRL1, pushes RBX1 from its WHB cradle (Figure 4.9b). It is useful to remind the reader that the primary functionality of the CRL as an E3 ligase comes from the RBX subunit. Conjugation of NEDD8 to CRLs appears to result in major conformational changes in the CRL C-terminal, leading to dramatic rotation of the WHB domain away from its original RBX-cradling position.

This appears to enhance flexibility of the RBX subunit which is free to mediate E2~Ub recruitment. Conformational changes in the CRL C-terminal also result in the inability of CAND1 to associate due to the disruption of the C-terminal binding site[175].

Structural basis of the Cullin 2 RING E3 ligase regulation by the COP9 signalosome

4.2 Abstract

Cullin-Ring E3 Ligases (CRLs) regulate a multitude of cellular pathways through specific substrate receptors. The COP9 signalosome (CSN) deactivates CRLs by removing NEDD8 from activated Cullins. Here we present structures of the neddylated and deneddylated CSN-CRL2 complexes by combining single-particle cryo-electron microscopy (cryo-EM) with chemical cross-linking mass spectrometry (XL-MS). These structures suggest a conserved mechanism of CSN activation, consisting of conformational clamping of the CRL2 substrate by CSN2/CSN4, release of the catalytic CSN5/CSN6 heterodimer and finally activation of the CSN5 deneddylation machinery. Using hydrogen-deuterium exchange (HDX)-MS we show that CRL2 activates CSN5/CSN6 in a neddylation-independent manner. The presence of NEDD8 is required to activate the CSN5 active site. Overall, by synergising cryo-EM with MS, we identify sensory regions of the CSN that mediate its stepwise activation and provide a framework for understanding the regulatory mechanism of other Cullin family members.

4.3 Introduction

Cullin-RING Ligases (CRLs) are modular, multi-subunit complexes that constitute a major class of ubiquitin E3 ligases[170,173]. CRLs coordinate the ubiquitination of substrates as either a signal for degradation via the 26S proteasome, or to alter the function of the target protein[170,209]. The CRL2 E3 ligase consists of a Cullin 2 (CUL2) scaffold in association with a catalytic RING-box protein (RBX1), with the substrate adaptors Elongin B (ELOB) and C (ELOC) at its N-terminal[198]. When associated with the von Hippel–Lindau (VHL) tumour suppressor substrate receptor, the CRL2 complex is the primary regulator of the Hypoxia Inducible Factor 1-α (HIF-1α) transcription factor[210,211]. Mutations in the interface between VHL, ELOB and ELOC can deactivate CRL2 leading to an accumulation of HIF-1α, which can in turn drive tumorigenesis through the over-activation of oncogenes[212]. Moreover, CRL2 has recently been identified as a potential target for small molecular inhibitors and PROTACs—a new class of cancer drugs that promote degradation of tumorigenic gene products[182,213,214]. These fascinating systems have been described in detail by a number of excellent reviews[170,173,209].

Activation of CRL2, in common with other members of the CRL family, involves a cascade of E1, E2 and E3 enzymes, which conjugate the ubiquitin-like protein NEDD8 (N8) to residue K689 on the CUL2 scaffold[215]. In its activated state, CRL2~N8 (the ~ stylisation denotes a covalent interaction) recruits the ubiquitin-conjugated E2 enzyme via the RING domain of RBX1[179]. Ubiquitination now takes place, covalently adding ubiquitin to the substrate molecule docked at the CRL2 N-terminal. The activity of CRL2 is negatively regulated by the 331 kDa Constitutive Photomorphogenesis 9 Signalosome (CSN) complex, frequently referred to as the COP9 signalosome complex[216-218]. The CSN was originally identified as consisting of eight subunits (designated as CSN1–8 by decreasing molecular weights of 57–22 kDa), and is organised in a splayed hand architecture, which has high sequence and

structural homology to the proteasome lid[216,217,219-221]. CSN1, 2, 3, 4, 7 and 8 are structurally homologous to each other and together contribute to the fingers of the splayed hand which arise from extended N-terminal α-helical repeats[219,221]. Each CSN1-8 subunit includes an extended C-terminal helix which associates together forming a C-terminal helical bundle. CSN5 and 6 are also closely related structurally and form a globular heterodimer located on the palm of the hand. CSN5 is responsible for the deneddylase activity of the CSN. A ninth subunit, CSNAP, has recently been identified, and is thought to play a role in stabilising the CSN complex[221].

Electron microscopy (EM) based structural analysis has provided important insights into the mechanism of CRL1 regulation by CSN[219,222]. CRL4A and CRL3 have also been observed to form such complexes[223]. However, despite intense interest, structural information of the CSN bound to CRLs remains limited to CRL1[179,219,222], CRL4A[223], and a low-resolution map of a dimeric CSN–CRL3~N8[223] complex. In particular, the analysis of the CSN–CRL4A~N8 complex[223] identified at least three major steps by which CRL~N8 is deneddylated by the CSN. In the first step, the extended N-terminal helical modules of CSN2 and CSN4 conformationally clamp the C-terminal domain of the CRL4A~N8 and RBX1[219,222,223]. The second step involves the release and consequent relocation of CSN5/CSN6 closer to NEDD8, brought about by disruption of the CSN4/CSN6 interface[223]. Disrupting the binding interface between CSN4/CSN6 through removal of the CSN6 insertion-2 loop (Ins-2), resulted in enhanced deneddylase activity[217], presumably due to more complete release of CSN5/CSN6. In the final step, the mobile CSN5 binds to NEDD8, leading to deneddylation via its JAB1/MPN/MOV34 (JAMM) metalloprotease domain[224]. The JAMM motif consists of H138, H140 and D151 zinc-coordinating residues, and residue E104 of the CSN5 insertion-1 loop (Ins-1)[217]. In apo-CSN, the Ins-1 loop occludes the CSN5 active site, auto-inhibiting the deneddylase[225]. Deneddylation is also severely diminished by a H138A point mutation in CSN5[217].

Surprisingly, the CSN can also form complexes with each of the Cullin 1-5 family members even without NEDD8[226]. Free CRLs such as CRL1 have been reported to readily bind and inhibit the CSN, albeit at relatively lower affinity than the neddylated CRL1[227]. While the exact role of CSN-CRL complexes remains unclear, it has been hypothesised that these complexes may function to regulate the cellular level of ubiquitin ligase activity of CRLs once they have been deneddylated, effectively sequestering E3 ligases from the intracellular environment[227].

Building on the existing knowledge of the CSN–CRL systems, here we pose the question: are similar structural changes to be found in other CSN–CRL complexes, and how does binding of neddylated CRLs lead to activation of the CSN5 catalytic site? To address this, we present structures of the CSN–CRL2~N8 complex, together with the structure of the CSN–CRL2 deneddylation product. We complement our cryo-EM analysis with chemical cross-linking mass spectrometry (XL-MS) allowing us to clarify the positions of particularly dynamic regions in the complexes. We use hydrogen-deuterium exchange mass spectrometry (HDX-MS) to interrogate the role of the CSN4/CSN6 interface in communicating CRL binding to the CSN5 active site. Overall, our structures of the CSN–CRL2~N8 and its deneddylation product, the CSN–CRL2, reveal the intricate conformational changes of CSN that lead to deneddylation of the CRL2.

4.4 Materials and Methods

4.4.1 Preparation and expression of bacmids

WT and catalytically reduced CSN^{5H138A} bacmids were a kind gift from Radoslav Enchev (The Francis Crick Institute, London)[222]. pcDNA3-myc3-CUL2 was a gift from Yue Xiong (Addgene plasmid #19892[228]). HA-VHL wt-pBabe-puro was a gift from William Kaelin (Addgene plasmid #19234[229]). RBX1, ELOB and ELOC were cloned from cDNA from the Mammalian Gene Collection (MGC) purchased from Dharmacon. To improve protein yield, an N-terminally truncated (1-53) natural isoform of VHL was also produced for use with mass spectrometry. Both isoforms of VHL were sub-cloned into a pET-52b(+) vector (Novagen) to add an N-terminal Strep-Tag II. Genes were assembled into pACEBac1 using I-CeuI/BstXI restriction sites via the MultiBac system[230]. RBX1 and CUL2 were assembled into one vector and ELOB, Strep II-VHL(ΔN) and ELOC into a second vector. Correct assembly was confirmed by sequencing of entire genes. DH10EmBacY cells were transformed with each assembly and blue/white selection was performed on L-agar plates containing 50 μg ml^{-1} kanamycin, 7 μg ml^{-1} gentamycin, 10 μg ml^{-1} tetracyclin, 100 μg ml^{-1} Bluo-Gal (Thermo Scientific) and 40 mg ml^{-1} IPTG. DH10MultiBac bacmid DNA was isolated from single white colonies. Recombinant baculoviruses were generated in Sf9 insect cells (a clonal isolate from Sf21, Life Technologies #11496015) using standard amplification procedures.

4.4.2 Expression and Purification of Recombinant CRL2

High Five Cells (BTI-TN-5B1-4, from embryonic tissue of the cabbage looper, *Trichoplusia ni*, Life Technologies #B85502) were co-infected with bacmids containing RBX1/CUL2 and ELOB/Strep-II VHL(ΔN)/ELOC and incubated at 27 °C and 130 rpm for 72 hours. Cells were harvested by centrifugation at 250xg for 10 mins at 4 °C before storage at -80 °C. Freeze-thawed pellets were resuspended in 50 mM Tris

pH 7.5, 150 mM NaCl, 2 mM DTT containing complete EDTA-free protease inhibitor tablets (Roche) and Benzonase (Sigma-Aldrich). Cells were lysed by sonication and clarified by centrifugation at 25,000 xg for 1 hour (Beckman JA-20 rotor). Supernatant was bound to a 3x 5 ml StrepTrap HP columns (GE Healthcare) in tandem, equilibrated with 50 mM Tris pH 7.5, 150 mM NaCl, 2 mM DTT. Protein was eluted by the addition of 2.5 mM *d*-desthiobiotin. The eluted peak fractions were concentrated to 2 ml and loaded onto a Superdex 200 16/600 (GE Healthcare) size exclusion column equilibrated with 50 mM HEPES pH 7.5, 150 mM NaCl, 1 mM TCEP. All CRL2 and CRL2~N8 samples used throughout were in 50 mM HEPES pH 7.5, 150 mM NaCl, 1 mM TCEP.

4.4.3 *In vitro* Neddylation of CRL2

APPBP1-Uba3, UbcH12 and Nedd8-His were purchased from (Enzo Life Sciences). The neddylation reaction was carried out for 10 min at 37 °C with 8 µM CRL2, 350 nM APPBP1-Uba3, 1.8 µM UbcH12 and 50 µM Nedd8 in a reaction buffer (50 mM HEPES pH 7.5 and 150 mM NaCl) supplemented with 1.25 mM ATP and 10 mM MgCl2. The reaction was quenched with 15 mM DTT and ice prior to loading onto a 1 ml StrepTrap HP column (GE Healthcare). CRL2~N8 was eluted with reaction buffer supplemented with 2.5 mM desthiobiotin and neddylation was confirmed by SDS-PAGE.

4.4.4 Expression and Purification of Recombinant CSN

High Five Cells were co-infected with the bacmids gifted by Radoslav Enchev (The Francis Crick Institute, London) and protein were expressed as described for CRL2, with an additional Ni-affinity step prior to gel filtration to exploit the His6-tag on the CSN5 subunit. For Strep-affinity chromatography 50 mM HEPES pH 7.5, 200 mM NaCl, 2 mM TCEP and 4% glycerol buffer was used, with the addition of 2.5 mM *d*-desthiobiotin for elution. For Ni-affinity using 2× HisTrap HP columns (GE Healthcare) in tandem, the same buffer was used, but protein was eluted by a 0–300

mM imidazole gradient across 45 ml. For size exclusion chromatography using a Superdex 200 16/600 (GE Healthcare), the column was equilibrated with 50 mM HEPES pH 7.5, 150 mM NaCl, 1 mM TCEP and 2% glycerol. All CSN and CSN^{WT} samples used throughout were in 50 mM HEPES pH 7.5, 150 mM NaCl, 1 mM DTT and 1% glycerol.

4.4.5 Cryo-EM of CSN-CRL2~N8

The CSN–CRL2~N8 complex was formed by incubation between CRL2~N8 (1.1× molar excess) and CSN at room temperature for 90 min. The preparation (~0.5 MDa) was subjected to size exclusion chromatography using a Superose 6 Increase 3.2/300 column (GE Healthcare), equilibrated in 15 mM HEPES pH 7.5, 100 mM NaCl, 0.5 mM DTT and 1% glycerol to reduce the contribution of apo components. Fractions from the leading edge of the peak were buffer exchanged into 15 mM HEPES pH 7.5 and 100 mM NaCl using PD SpinTrap G-25 columns (GE Healthcare) before initial assessment by negative stain EM. Cryo-grids were prepared using a Vitrobot (FEI). A cryo-EM dataset was collected beamline M02 from Quantifoil grids with an extra carbon layer at the Electron Bio-Imaging Centre (eBIC - Diamond Light Source, UK) on a Titan Krios 300 kV with Gatan K2 detector (M02), Å pix^{-1} = 1.06. Movies of 25 frames (dose = 1.85 e $Å^{-2}$) were motion corrected in RELION[231] (2.0) using MOTIONCOR2[232] (01-30-2017) and subsequent CTF estimation of micrographs was performed using CTFFIND4[233] (4.1.5) (Supplementary Figure 6.18). Auto-picking selected ~317,000 particles from ~3100 micrographs. Particles were subjected to reference free 2D classification to assess data quality and to remove contaminants selected by auto-picking. This process reduced the particle number to ~250,000. Following particle selection through 2D classification, particles were divided into 15 3D classes. Three of these classes (~69,000 particles) were selected for further classification and processing, as described in Supplementary Figure 6.19.

4.4.6 Cryo-EM of CSN-CRL2

The CSN–CRL2 complex was formed by incubation between CRL2 (1.1× molar excess) and CSN at room temperature for 90 min. Samples were loaded onto a 5–50% glycerol GraFix[234] gradient containing 0–0.2% glutaraldehyde and ultracentrifuged at 86,000 ×*g* for 24 h at 4 °C. Gradients were manually fractionated and the resultant aliquots assessed by SDS-PAGE to determine the extent of cross-linking. Fractions were also assessed using negative stain EM in-house. In order to reduce the glycerol content of samples for cryo-EM, fractions containing the desired complex (as determined by negative stain EM) were pooled together and gel filtered into 15 mM HEPES pH 7.5, 100 mM NaCl and 0.5 mM DTT using a Superose 6 Increase 10/300 GL column (GE Healthcare). Fractions were again assessed by negative stain before preparing cryo-grids using a Vitrobot (FEI). A cryo-EM dataset was collected at the Electron Bio-Imaging Centre (eBIC - Diamond Light Source, UK) on a Titan Krios 300 kV with Gatan K2 detector (M02), with a sampling of 1.047 Å pix^{-1}. Movies of 85 frames (dose = 1.0 e $Å^{-2}$) were motion corrected in RELION[235] (3.0) using MOTIONCOR2[232] (01-30-2017) and subsequent CTF estimation of micrographs was performed using CTFFIND4[233] (4.1.5). Auto-picking selected ~309,000 particles from ~6800 micrographs. Particles were subjected to reference free 2D classification to assess data quality and to remove contaminants selected by auto-picking. This process reduced the particle number to ~208,000. Following particle selection through 2D classification, particles were divided into six 3D classes first with alignment, then subsequently without alignment with a mask around CSN5/CSN6 in order to perform focused classification on this area. The map that showed the greatest recovery of detail for CSN5/CSN6 (Supplementary Figure 6.28)[215] was then subjected to 3D auto refinement and post-processing. Local resolution was estimated using ResMap[236] as part of the RELION wrapper.

4.4.7 Band-shift assays

In order to test the activity of the CSN and CSN^{WT}, 3 µg of each complex was separately incubated with 3 µg of CRL2~N8 at 37 °C for 0, 15, 30, 45 and 60 s. Deneddylation was quenched through rapid denaturation by the addition of NuPAGE lithium dodecyl sulphate sample buffer (Thermo Fisher) and placing on a heat block, pre-heated to 90 °C. Samples were analysed by SDS-PAGE. Deneddylation in samples with the CSN^{WT} were confirmed by a band shift the gel band corresponding to CUL2, by comparison with CRL2~N8 and CRL2 controls.

4.4.8 Homology modelling of the CRL2

Homology modelling of the CRL2 was necessary due to a combination of missing domain structure for the CUL2 Winged-Helix A (WHA) and VHL subunit in the only crystal structure of CRL2 (PDB 5N4W). Homology modelling was performed through two stages: first, generating a CRL2 structure with a correct WHA domain, and second, generating the complete CRL2 intact with the VHL-ELOB-ELOC adaptor complex. In the first stage, we performed structural alignment of the CRL2 (5N4W) and CRL1 (1LDJ) structures in PyMOL (2.0.6). Using MODELLER[237], a single model of the CRL2 was generated using the slow molecular refinement option of MODELLER. The model was manually evaluated for correct fold, including the correct positioning of residues (such as the CUL2 K689 NEDD8-acceptor site) already present in the 5N4W crystal structure. In the second stage, we aligned the CRL2 model with the VHL-ELOB-ELOC-CUL2 fragment (4WQO) to generate a template for homology modelling. The isoform 3 of VHL (missing residues 1–53) was used for modelling to maintain consistency with the experimental construct. Again, a single model of the CRL2 (with VHL/ELOB/ELOC) was generated using the slow molecular refinement option of MODELLER. The final model of the CRL2 shows a RMSD of 3.6 Å when compared with the initial crystal structure (5N4W) but includes a complete WHA domain and VHL subunit.

4.4.9 Model fitting of EM maps

All models were fitted using CSN subunits sourced from the 4D10 crystal structure (chains A–H) and the CRL2 (VHL-ELOB-ELOC) structure generated and described in section 4.4.8. We performed map fitting first by performing rigid body fitting of the CSN and CRL2 subunits to each map in Chimera[238] (1.13.1rc) then using the Molecular Dynamics Flexible Fitting (MDFF)[239] (0.5) feature of NAMD[240] for positional refinement. In the rigid body fitting step, elongated subunits such as CSN2, CSN4 and CUL2 were first dissected into smaller rigid bodies to permit better fitting into their densities. Following map fitting of all CSN and CRL2 subunits, we then converted the structures into MDFF-compatible topology files using the protein structure file builder function of VMD[62] (1.9.3). MDFF was performed in two steps: an initial energy minimisation step (scaling factor = 0.3 for 50,000 steps) which coerced each subunit into their map densities, and a second equilibration run (scaling factor = 10 for 200,000 steps) which applied molecular dynamics to produce structurally and energetically realistic structures. Secondary structure, *cis*-peptide and chirality characteristics of the initial models were calculated and enforced throughout each step to avoid a loss of internal structure for each subunit. For each of the CSN–CRL2~N8 and CSN–CRL2 structures, subunits/domains which lacked clear density were not included to avoid interference with the fitting of other subunits. These were the WHB domain (CUL2 residues 656–745) for all maps, NEDD8 in all NEDD8-including maps, and VHL in the CSN–CRL2 map. The cross-correlation coefficient which calculates the degree of overlap between the cryo-EM map and a simulated map of the same resolution from the atomic model, are reported for each model in Table 4.3 of the manuscript.

4.4.10 Native mass spectrometry

All spectral data were collected using a SYNAPT G2-Si (Waters Corp., Manchester, UK) high-definition mass spectrometer and samples were ionised using a NanoLockSpray™ dual electrospray inlet source (Waters Corporation) run with

positive polarity in sensitivity mode. Capillaries were pulled using a Flaming/Brown P-97 micropipette puller (Sutter Instrument) and coated with Au:Pd (80:20) using a sputter coater (Quorum Q150RS). The following mass spectrometer settings were used: capillary voltage 1.60–1.75 kV, sampling cone of 75–150 V, source temperature of 20 °C, desolvation temperature 150 °C and collision energy of 25–75 eV. Gas pressures were: source 9.3×10^{-3} mbar, trap 3.3×10^{-2} mbar, helium cell 3.4 mbar, drift tube 2.6 mbar, transfer 3.1×10^{-2} mbar and time-of-flight 5.7×10^{-7} mbar.

A 1:1 ratio of CSN:CRL2~N8 and CSN^{WT}:CRL2 at a 5–15 µM concentration were pre-incubated for 1 h prior to buffer exchange. Pre-incubated protein samples were buffer exchanged and desalted using Vivaspin 500 (30 kDa MWCO) centrifugal concentrators (Sartorius) into pH 7.5 150 mM ammonium acetate (four wash steps). All spectra were analysed using MassLynx (4.1, Waters Corp.).

4.4.11 Hydrogen deuterium exchange mass spectrometry

HDX-MS experiments were performed on a Synapt G2-Si HDMS coupled to an Acquity UPLC M-Class system with HDX and automation (Waters Corporation, Manchester, UK). Data were collected in positive polarity in sensitivity mode and calibrated using sodium iodide. The following mass spectrometer settings were used: capillary voltage of 3 kV, sampling cone of 100 V, source temperature of 80 °C and desolvation temperature of 150 °C. Acquisition mass range was set to 50–2000 Da. Gas pressures were: source 6.6×10^{-3} mbar, trap 2.9×10^{-2} mbar, helium cell 4.5 mbar, drift tube 3.1 mbar, transfer 2.8×10^{-2} mbar and time-of-flight 8.2×10^{-7} mbar.

Protein samples were prepared at a concentration of 7.5 µM. Isotope labelling was initiated by diluting 5 µl of each protein sample into 95 µl of buffer L (10 mM potassium phosphate in D_2O pD 6.6). The protein was incubated at various time points (0.25, 5 and 30 min) and then quenched in ice cold buffer Q (100 mM potassium phosphate, brought to pH 2.3 with formic acid (FA)) before being digested online with a Waters Enzymate BEH pepsin column at 20 °C. The same procedure was used for undeuterated control, with the labelling buffer being replaced by buffer

E (10 mM potassium phosphate in H_2O pH 7.0). The peptides were trapped on a Waters BEH C18 VanGuard pre-column for 3 min at a flow rate of 200 µl min^{-1} in buffer A (0.1% FA ~pH 2.5) before being applied to a Waters BEH C18 analytical column. Peptides were eluted over 7 min with a linear gradient of buffer B (8–40% gradient of 0.1% FA in acetonitrile) at a flow rate of 40 µl min^{-1} with a runtime of 11 min. All trapping and chromatography were performed at 0.5 °C to minimise back exchange. MSE data were acquired with a 25–45 eV transfer collision energy ramp for high-energy acquisition of product ions. Leucine Enkephalin (LeuEnk-Sigma) was used as a lock mass for mass accuracy correction and the MS was calibrated with sodium iodide. The online Enzymate pepsin column was washed with pepsin wash (1.5 M Gu-HCl, 4% MeOH and 0.8% FA) recommended by the manufacturer and a blank run using the pepsin wash was performed between each sample to prevent significant peptide carry-over from the pepsin column. Optimised peptide identification and peptide coverage for all samples was performed from undeuterated controls (three–four replicates). All deuterium time points were performed in triplicate on different samples on distinct samples.

Sequence identification was made from MSE data from the undeuterated samples using the Waters ProteinLynx Global Server 2.5.1 (PLGS). Processing parameters of PLGS were set to: lock mass for charge 1 of 556.2771 Da e^{-1}, lock mass window of 0.4 Da, low energy threshold of 135.0 counts, elevated energy threshold of 30.0 counts and intensity threshold of 750 counts. Workflow parameters were: peptide and fragment mass tolerance set to automatic, minimum fragment ion matches per peptide set to 1, minimum fragment ion matches per protein set to 7, minimum peptide matches per protein set to 3, primary digest reagent set to non-specific, number of missed cleavages 0, false discovery rate of 100%.

The output peptides were filtered using DynamX (3.0) using the following filtering parameters: minimum intensity of 2500, minimum and maximum peptide sequence length of 5 and 30, respectively, minimum MS/MS products of 3, minimum products per amino acid of 0.1, and a minimum peptide score of 5. In addition, all the spectra

were visually examined and only those with high signal to noise ratios were used for HDX-MS analysis. The amount of relative deuterium uptake for each peptide was determined using DynamX (3.0) and are not corrected for back exchange. State (listing the deuterium uptake per-peptide, per-timepoint and per-experimental state) and difference (listing the difference in deuterium uptake between identical peptides of two states compared for each timepoint), were exported from DynamX. These files were input to Deuteros[138] (1.0.8) which format the differential data into the Woods Plot format. Statistical filtering of peptides is then performed to identify those which exhibit significant uptake differences between the two states. Deuteros applies a blanket confidence interval (specifically, the 98% confidence interval given as 0 ± DU where DU is the Deuterium uptake threshold in Daltons) across all peptides of each timepoint. Significant peptides for each timepoint are then exported into a formatting script which is used to project the filtered data onto a 3D model of the protein (Supplementary Figure 6.36-Supplementary Figure 6.37).

4.4.12 PLIMSTEX for CSN-CRL2 complexes

In addition to performing traditional time-resolved HDX-MS experiments, we also performed PLIMSTEX (protein–ligand interactions by MS, titrations and HDX) measured on combinations of CSN/CSN^{WT} and CRL2/CRL2~N8 to derive dissociation constants (Kd). We performed PLIMSTEX for three complexes: (1) CSN–CRL2~N8, (2) CSN–CRL2 and (3) CSN^{WT}–CRL2 complexes. The final concentration of CSN or CSN^{WT} was fixed at 250 nM. CRL2 or CRL2~N8 were titrated at either 1:0, 1:0.1, 1:0.5, 1:1, 1:2 or 1:0, 1:0.1, 1:0.5, 1:1, 1:5 molar ratios of CSN:CRL2 (0–1250 nM). The sample setup for each of the three complexes consisted of five undeuterated references of CSN or CSN^{WT}, three samples of deuterated CSN or CSN^{WT}, followed by three of each of the above molar ratios of CSN:CRL2. Datasets for CSN–CRL2~N8, CSN–CRL2 or CSN^{WT}–CRL2 underwent labelling for either 15 sec or 2 min depending on which exposure time yielded better deuterium uptake differences as a function of [CRL2] or [CRL2~N8]. HDX-MS data acquisition and data analysis using PLGS and DynamX were

performed as above. Kd values for each complex were derived using a MathCAD worksheet (v14, Parametric Technology Corp., Needham, USA) kindly provided by Michael Gross (Washington University in St. Louis, USA). Briefly, the deuterium uptake of each peptide as a function of increasing [CRL2] or [CRL2~N8] were fitted using a 1:1 binding model, optimising for three parameters: D0 (initial deuterium uptake in the absence of [CRL2] or [CRL2~N8]), ΔD (maximum decrease in deuterium uptake observed) and β (where β is equal to the association constant, *K*a, for 1:1 binding). The quality of the fit is calculated as the root mean square of the residuals between experimental datapoints and the model values. The PLIMSTEX data fitting process has been documented in much greater detail elsewhere[67,241,242].

4.4.13 Chemical cross-linking mass spectrometry

Twenty microlitres of ~20 µM CSN–CRL2~N8, CSN^{WT}–CRL2 and apo-CSN^{WT} were each incubated with 1–5 mM BS3 cross-linker for 1 h at 25 °C and 350 r.p.m. in a thermomixer. After cross-linking, complexes were (i) separated by gel electrophoresis (NuPAGE) followed by in-gel digestion (CSN^{WT}–CRL2, CSN^{WT} and CSN–CRL2~N8) or (ii) digested in solution (CSN^{WT}–CRL2) and generated peptides were pre-fractionated by gel filtration (CSN–CRL2~N8). Gel electrophoresis was performed using the NuPAGE system according to manufacturer's protocols. In-gel digestion was performed as described before[243]. Digestion in solution was performed in the presence of RapiGest (Waters) according to manufacturer's protocols. For gel filtration, peptides were dissolved in 30% acetonitrile (MeCN), 0.1% trifluoroacetic acid and separated on a Superdex Peptide PC 3.2/30 column (GE Healthcare) at a flow rate of 50 µl min^{-1}.

4.4.14 Mass spectrometry for XL-MS

Peptides were dissolved in 2% MeCN, 0.1% FA and separated by nano-flow liquid chromatography (Dionex UltiMate 3000 RSLC, Thermo Scientific; mobile phase A: 0.1% (v/v) FA; mobile phase B: 80% (v/v) MeCN, 0.08% (v/v) FA). Peptides were loaded

onto a trap column (µ-Pre-column, C18, 100 µm I.D., particle size 5 µm; Thermo Scientific) and separated with a flow rate of 300 nl min^{-1} on an analytical C18 capillary column (Acclaim PepMap 100, C18, 75 µm I.D., particle size 3 µm, 50 cm; Thermo Scientific), with a gradient of 4–90% (v/v) mobile phase B over 66 min. Separated peptides were directly eluted into a Q Exactive Plus hybrid quadrupole-Orbitrap (CRL2 and CRL2~N8) or an Orbitrap Fusion Tribrid Mass Spectrometer (CSN–CRL2~N8) (Thermo Scientific).

Typical mass spectrometric conditions for the Q Exactive Plus were: spray voltage of 1.6–2.1 kV; capillary temperature of 250 °C; normalised collision energy of 30%, activation Q of 0.25. The mass spectrometer was operated in data-dependent mode. Survey full scan MS spectra were acquired in the Orbitrap from 350–1600 or 2000 *m/z* with a resolution of 70,000 at an automatic gain control (AGC) target of 3 × 106. The top 20 most intense ions were selected for Higher Energy Collisional Dissociation (HCD) MS/MS fragmentation in the Orbitrap (isolation window, 1.5 or 1.6 *m/z*). MS/MS spectra were acquired at a resolution of 17,500 at an AGC target of 1 × 105 or 5 × 104. Previously selected ions within previous 30 s were dynamically excluded for 30 s. Only ions with charge states 2–7+ were selected. Singly charged ions as well as ions with unrecognised charge state were excluded. Internal calibration of the Orbitrap was performed using the lock mass option (lock mass: *m/z* 445.120025)[244].

Typical mass spectrometric conditions for the Orbitrap Fusion were: spray voltage of 2.5 kV; capillary temperature of 275 °C; collision energy of 30% and activation Q of 0.25. The mass spectrometer was operated in data-dependent mode. Survey full scan MS spectra were acquired in the Orbitrap from 500–1700 or 2000 *m/z* with a resolution of 120,000 at an AGC target of 5 × 104. The most intense ions were selected for HCD MS/MS fragmentation in the Orbitrap (3 s cycle time with a maximum injection time of 128 ms; isolation window, 1.6 *m/z*). MS/MS spectra were acquired at a resolution of 30,000 at an AGC target of 5 × 104. Previously selected ions within previous 30 s were dynamically excluded for 20 s. Only ions with charge states 3–8+ were selected. Singly and doubly charged ions as well as ions with unrecognised

charge state were excluded. Internal calibration of the Orbitrap was performed using the lock mass option (lock mass: *m/z* 445.120025)[244].

4.4.15 Data analysis for XL-MS

Raw files were converted into Mascot generic format (mgf) files using pXtract[f]. Mgf's were searched against a reduced database containing CSN and CRL2 proteins using pLink 1.0 search engine. Search parameters were: instrument spectra, HCD; enzyme, trypsin; max missed cleavage sites, 3; variable modifications, oxidation (methionine) and carbamidomethylation (cysteine); cross-linker, BS3; min peptide length, 4; max peptide length, 100; min peptide mass, 400 Da; max peptide mass, 10,000 Da; false discovery rate, 1%. Potential cross-linked dipeptides were evaluated by their spectral quality. Circular network plots were generated using the XVis[g,61] webserver.

4.4.16 XL-MS guided placement of the WHB, NEDD8 and VHL subunits

Cross-links determined from XL-MS for the CSN-CRL2~N8 and CSN-CRL2 complexes were used to clarify the position of the WHB, NEDD8 and VHL subunits which lacked clear density in our cryo-EM maps. We performed XL-guided placement using the Integrative Modelling Platform (IMP)[108] (2.6), using as input, the map-fitted models of the CSN-CRL2~N8 and CSN-CRL2. The CSN-CRL2~N8 model post-map fitting, included all subunits except the WHB and NEDD8. Similarly, the CSN-CRL2 model included all subunits except the WHB and VHL. Separately, the subunits of each complex were initialized as coarse-grained bead models, representing each residue as a single bead. The WHB domain (Cullin-2 residues 656-745) and VHL was sourced from the homology model of the CRL2 (detailed in the section Homology Modelling of the CRL2). NEDD8 was sourced from the crystal structure of neddylated CRL5

f http://pfind.ict.ac.cn/software/pXtract/index.html

g https://xvis.genzentrum.lmu.de/login.php

(3DQV). WHB, NEDD8 and VHL were set as mobile rigid bodies, while all other subunits were kept stationary.

Our modelling procedure utilised two types of cross-links. The first type are pseudo-cross-links that maintain the correct topology of the complex: a single pseudo-cross-link between $CUL2^{T655}$ to WHB^{T656} of 5 Å to mimic a covalent bond, and connections between VHL-ELOB, VHL-ELOC and VHL-CUL2 to maintain integrity of the VHL-ELOB-ELOC adaptor complex and its interface with CUL2. A single pseudo-cross-link of 10 Å was used to mimic the isopeptide bond of WHB^{K689}~$N8^{G76}$ (7.5 Å lysine side chain + ~3 Å glycine C-terminus). The second type are cross-links determined experimentally between WHB, NEDD8 and VHL with its surrounding subunits which utilised a distance threshold of 35 Å (two lysine side chains at 15 Å, BS3 linker length at 10 Å, plus 10 Å for flexibility). IMP was parametrised to perform 1000 iterations, with each iteration randomly moving WHB, NEDD8 and VHL relative to the stationary CSN and CRL subunits. IMP parameters used were num_mc_steps = 10, rb_max_trans = 2, rb_max_rot = 0.1, bead_max_trans = 0.5 and excluded volume restraint resolution of 20. The single best model was evaluated by projecting all cross-links for the complex onto the structure and confirming that all distances were below the 35 Å distance threshold. A table of cross-links can be found in Supplementary Data 6.1 for CSN–CRL2~N8 and Supplementary Data 6.2 for CSN^{WT}–CRL2. The script used can be found in Supplementary Note 6.2.

4.5 Results

4.5.1 Cryo-EM structures of the CSN–CRL2~N8 complex

To study the molecular interactions between neddylated CRL2 (CRL2~N8) and the CSN, we performed single-particle cryo-EM to resolve a structure of the assembled CSN–CRL2~N8 (referred to as the holocomplex) (Supplementary Figure 6.18-Supplementary Figure 6.19). The H138A mutation in the catalytic site of CSN5 subunit makes it possible to assemble CSN–CRL2~N8 complexes in which NEDD8 remains covalently attached over the time scale of the experiment[219,222]. To justify the use of the H138A point mutation, we performed a band-shift assay comparing the deneddylation activity of the CSN^{WT} and CSN^{5H138A} mutant enzymes (Supplementary Figure 6.20). As expected, CSN^{WT} rapidly cleaves NEDD8 from CRL2~N8 within seconds of incubation, while little to no deneddylation activity is seen from the CSN^{5H138A} complex (Supplementary Figure 6.20). This mutant form of CSN was used throughout the work described below, unless otherwise specified.

Using 3D classification, we were able to generate maps of three different structures: (a) a holocomplex map at 8.2 Å, (b) a map of the complex with little or no density for VHL at 8.0 Å, and (c) a map of the complex with little or no density for CSN5/CSN6/VHL at 6.5 Å (Supplementary Figure 6.19, Supplementary Figure 6.21-Supplementary Figure 6.22). The two partial complexes likely arise from compositional heterogeneity in the original samples from which the structural analysis has succeeded in isolating subpopulations. To verify the existence of subcomplexes, we subjected the CSN–CRL2~N8 to native MS (Supplementary Figure 6.23). In line with subpopulation observations from our cryo-EM, we identified subcomplexes of the CSN-CRL2~N8 missing the CSN5, VHL, ELOB or ELOC subunits. These observations support the notion that similar levels of heterogeneity observed in the cryo-EM analysis of other CSN-CRL complexes, CSN-CRL1 and CSN-CRL4A[222,223]

is also likely to arise from variable subunit composition and may be ubiquitous to all CSN-CRL complexes.

Next, we fitted into each map, the highest resolution crystallographic structure of the CSN (PDB 4D10)[217] and a homology model of the CRL2 (including the VHL-ELOB-ELOC) using molecular dynamics flexible fitting[239] (MDFF; Table 4.1, 4.4.9 Model fitting of EM maps). In the model of the holocomplex (8.2 Å), the main interactions occur between the C-terminal end of CUL2 and the extended N-terminal helical repeats of CSN2 and CSN4 (Figure 4.10a-b). Compared to their conformation in the highest resolution apo-CSN crystal structure[217] (PDB 4D10), CSN2 and CSN4 are moved by 30 and 51 Å, respectively, towards CUL2 (Figure 4.10e, Supplementary Movie 6.1). We additionally reviewed the conformations of CSN2 and CSN4 in all nine available intact structures of the apo-CSN (PDBs 4D10, 4D18 and 4WSN; Supplementary Figure 6.24). We compared each of the apo-CSN structures with our holocomplexes and with other fitted models of the CSN in complex with CRL1~N8 and CRL4A~N8. Significant structural variation in CSN2 and CSN4 is observed within these apo-CSN structures, but in all cases, their conformations were substantially different to any found in the holocomplexes. In each of our CSN–CRL2 structures, the clamping motion of CSN2 and CSN4 is a swinging rotation about hinges located close to the CSN2 and CSN4 winged helix domains. For CSN2 this is coupled with an additional rotation about the axis of the superhelix formed from its N-terminal helical repeats. In the case of CSN4 the movement is coupled to the detachment of CSN4 from the Ins-2 loop of CSN6 by ~30 Å and leads to an ~12 Å shift in CSN5 (Figure 4.10c, f). Only minor conformational changes were found in the CSN1, CSN3, CSN7B or CSN8 subunits. In the CRL2~N8 moiety, a number of relatively small rearrangements of CUL2, RBX1, ELOB, ELOC and VHL subunits were observed compared to its crystal structure[182] (Supplementary Figure 6.25a-b).

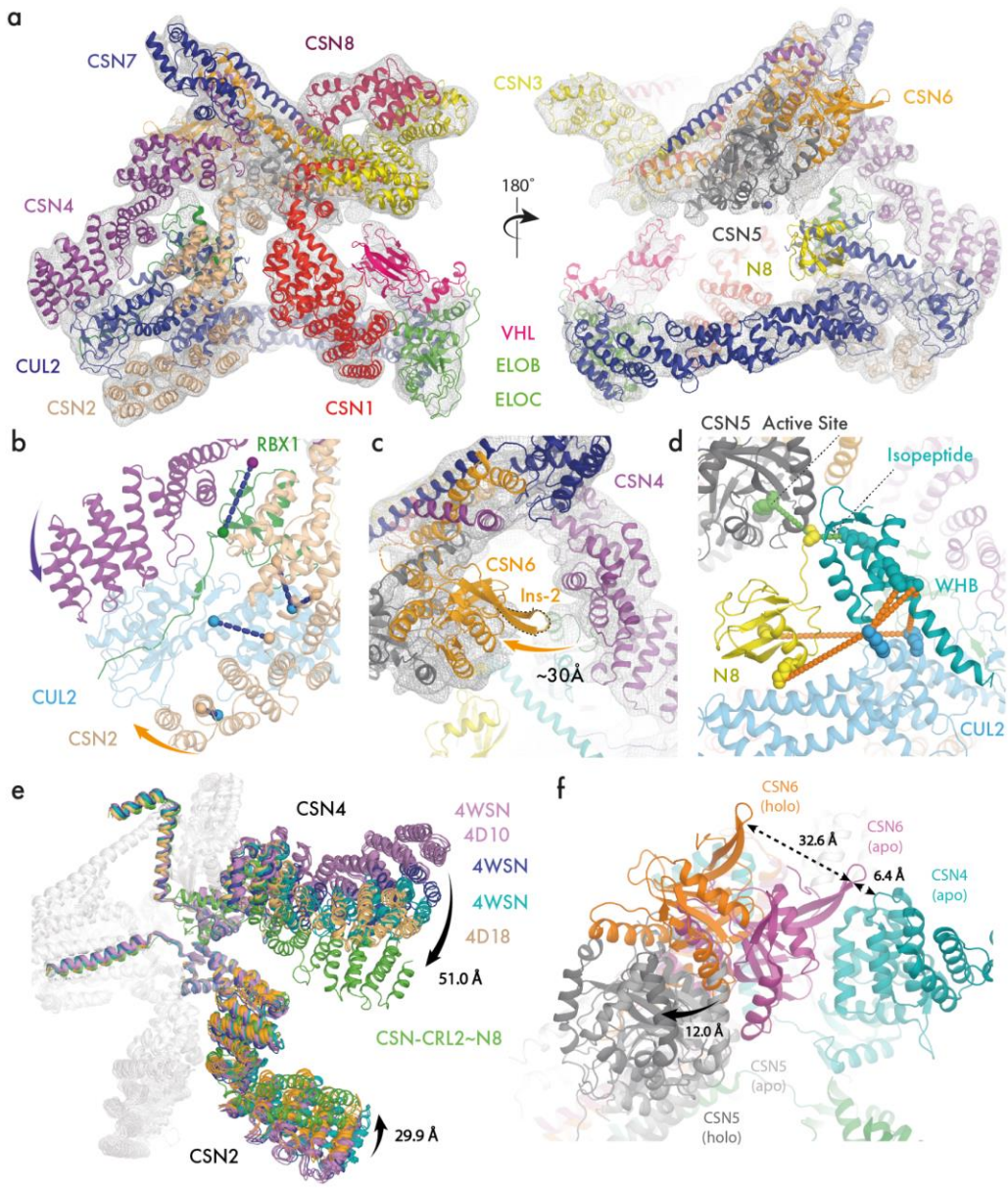


Figure 4.10. Structures and interactions of the CSN–CRL2~N8 complex. (a) The molecular model of CSN–CRL2~N8 fitted into cryo-EM density (8.2 Å resolution) from front and back views. (b) Conformational clamping of CRL2~N8 by CSN2 and CSN4. Cross-links shown are between CSN4-RBX1 ($CSN4^{K200}$-$RBX1^{K105}$, purple-green spheres), and four between CSN2–CUL2 ($CSN2^{K157}$-$CUL2^{K489}$, $CSN2^{K263}$–$CUL2^{K462}$, $CSN2^{K225}$–$CUL2^{K462}$, $CSN2^{K64}$–$CUL2^{K404}$, beige-blue spheres). (c) View showing ~30 Å separation of CSN6 Ins-2 loop from CSN4 following CRL2~N8 binding. (d) Modelled position of WHB~N8 using cross-links of the CSN–CRL2~N8. Large scale conformational changes (e) between CSN2 and CSN4 in all apo-CSN crystal structures (PDB 4WSN, 4D10 and 4D18) and CSN–CRL2~N8 (f) CSN5/CSN6 (PDB 4D10) upon binding of CRL2~N8 (holo). Subunits of the CSN–CRL2~N8 were compared with the apo-CSN crystal structure (PDB 4D10) following structural alignment. The structure of the CRL2~N8 has been hidden for clarity.

In our EM maps and the other published CSN-CRL structures[219,222,223], the exact position of NEDD8 and the CUL2 Winged-Helix B (WHB) domain were difficult to determine. To address this limitation, we carried out XL-MS experiments on the CSN-CRL2~N8 complex using the bis(sulfosuccinimidyl)suberate (BS3) cross-linker which targets lysine sidechains (4.4 Materials and Methods). We identified a total of 24 inter- and 60 intra-protein cross-links (Supplementary Data 6.1, Supplementary Figure 6.26a-b). To generate a model of the CSN–CRL2~N8, we performed cross-link guided modelling which allows the placement of the WHB, NEDD8 and VHL subunits using identified cross-links from XL-MS (4.4 Materials and Methods). We imposed a cross-link distance threshold of 35 Å which takes into account the length of two lysine side chains (15 Å), the BS3 cross-linker length (10 Å) and an extra 10 Å to allow for domain-level flexibility (4.4 Materials and Methods). Our model of the CSN–CRL2~N8 satisfies all cross-link distances (Supplementary Figure 6.26c). Three cross-links between CUL2-WHB ($CUL2^{K382}$-WHB^{K720}, $CUL2^{K382}$-WHB^{K677} and $CUL2^{K433}$-WHB^{K677}) were used for the positioning of the WHB domain (Supplementary Figure 6.26d, red text). A further two cross-links between CUL2 and NEDD8 ($CUL2^{K382}$-$N8^{K33}$ and $CUL2^{K433}$-$N8^{K6}$) allowed the positioning of NEDD8 near CSN5 (Figure 4.10d, Supplementary Figure 6.26d, green text). In this conformation, the isopeptide bond of NEDD8 is juxtaposed to the CSN5 active site. For the isopeptide bond of NEDD8 to reach the CSN5 active site, the WHB domain must be extended from its crystallographic conformation towards the CSN5 by 19 Å (Supplementary Figure 6.26e).

Table 4.3. Cryo-EM data collection, refinement and validation statistics.

	CSN-CRL2~N8	CSN$^{-CSN5/6-}$CRL2^{-VHL}~N8	CSN-CRL2^{-VHL}~N8 (CSN5/6 REFINED)	CSN-CRL2~N8 (VHL REFINED)	CSN-CRL2
EMDB	4739	4744	4742	4736	4741
PDB ID	6R7F	6R7N	6R7I	6R6H	6R7H
SUBUNITS MISSING	None	CSN5/6, VHL	VHL	None	None
DATA COLLECTION AND PROCESSING					
MAGNIFICATION	47,170	47,170	47,170	47,170	47,755
VOLTAGE (kV)	300	300	300	300	300
ELECTRON EXPOSURE (eÅ^{-2})	45	45	45	45	83
DEFOCUS RANGE (mm)	1.8-3.0	1.8-3.0	1.8-3.0	1.8-3.0	1.8-3.0
PIXEL SIZE (Å)	1.060	1.060	1.060	1.060	1.047
SYMMETRY IMPOSED	C1	C1	C1	C1	C1
INITIAL NUMBER OF PARTICLES	316,921	316,921	316,921	316,921	308,936
FINAL NUMBER OF PARTICLES	20,055	22,471	24,552	24,049	17,191
MAP RESOLUTION (Å)	8.2	6.5	8.0	8.4	8.8
FSC THRESHOLD	0.143	0.143	0.143	0.143	0.143
MAP RESOLUTION RANGE (Å)	6.7-22.7	5.9-8.0	5.9-11.5	6.3-17.4	5.6-16.8
REFINEMENT					
INITIAL PDB USED	4D10, 5N4W, 4WQO, 3DQV	4D10, 5N4W, 4WQO	4D10, 5N4W, 4WQO	4D10, 5N4W, 4WQO	4D10, 5N4W, 4WQO
NUMBER OF RESIDUES	3727	2974	3578	3716	3251

4.5.2 Structure of the deneddylated CSN–CRL2 complex

Having determined the structure of the CSN-CRL2~N8 complex, we next sought to detail any conformational differences in the deneddylated CSN-CRL2. The affinity of CSN for the non-neddylated CRLs is significantly lower than for the neddylated

forms, limiting the yield of the desired product[222]. To stabilise the formation of a complex between CSN and CRL2, we employed GraFix[234] (4.4 Materials and Methods) prior to cryo-EM. Moreover, native MS confirmed the formation of CSN–CRL2 complex (Supplementary Figure 6.27). We next resolved a cryo-EM map of the CSN–CRL2 to 8.8 Å resolution (Supplementary Figure 6.28-Supplementary Figure 6.29). Although the resolution of the CSN–CRL2 map is similar to that for CSN–CRL2~N8 holocomplex, we only observed partial density for VHL and CSN4. Next using the same procedure as for the neddylated CSN–CRL2~N8 complex, we fitted CSN and CRL2 subunits into the density map of the CSN–CRL2 complex (4.4 Materials and Methods). We then utilised cross-link guided modelling to establish the position of the WHB which lacked clear density, similar to the neddylated holocomplex (Supplementary Figure 6.30, Supplementary Data 6.2, 4.4 Materials and Methods).

To determine whether the lack of density for the CSN4 may be due to flexibility of the CSN4 N-terminal domain, we carried out XL-MS for the apo-CSN complex (Supplementary Figure 6.31, Supplementary Data 6.3). A total of 18 cross-links were identified involving CSN4: 12 of which were intra-CSN4 cross-links, 4 between CSN2-CSN4 and 2 between CSN6–CSN4. We measured the distances of these cross-links on the structure of the apo-CSN (PDB 4D10) and on our cryo-EM fitted model of the CSN–CRL2 (which represented the CSN4 in a lowered conformation; Supplementary Figure 6.31). Applying a 35 Å distance threshold, we identified cross-links which were exclusively satisfied in each of the two conformations of CSN4 (Supplementary Figure 6.31, 4.4 Materials and Methods). The presence of exclusively satisfied cross-links indicates that both conformations have been sampled experimentally and suggest that the apo CSN4 can wave between the two conformations represented by the crystal structure and CRL2-bound structures, even in the absence of the CRL2 substrate.

To evaluate any local changes across the CSN–CRL2~N8 in the absence of NEDD8, we aligned the cryo-EM models of neddylated and deneddylated holocomplexes using the C-terminal helical bundle as a reference point (Supplementary Figure

6.32a-b). We systematically compared the conformations of each subunit (Supplementary Figure 6.32c-j, Supplementary Movie 6.1). Compared with its structure in the CSN–CRL2~N8, the N-terminal helices of CSN2 are shifted by ~21 Å towards the CUL2 C-terminal domain (Figure 4.11). This change in CSN2 in the absence of NEDD8, leads to a structural difference in CUL2 which rotates upwards towards the rest of the CSN by 20 Å (Figure 4.11c). The position adopted by CUL2 in the deneddylated holocomplex, places ELOB closer to CSN1, forming a CSN1–ELOB interface (Figure 4.11d). An interface between CSN1–ELOB can also be seen in the partial structures of CSN–CRL2~N8 missing VHL and CSN5/CSN6 (Supplementary Figure 6.19, Supplementary Figure 6.33). The dissociation of VHL and CSN5/CSN6 causes the N-terminus of CUL2 to shift downwards, away from the CSN (Supplementary Figure 6.33). The plasticity of the CSN results in changes in the C-terminus of CUL2, which is clamped between CSN2 and CSN4, in order to accommodate this shift. The formation of the CSN1-ELOB interface appears to arise as result of this movement in both the deneddylated structure and the incomplete CSN-CRL2~N8 structures. Interactions between substrate adaptor complexes and CSN1 have similarly been reported for the CSN-CRL1~N8[219] and CSN-CRL4A~N8[223]. RBX1 remains clamped between CSN2 and CSN4 (Figure 4.11e).

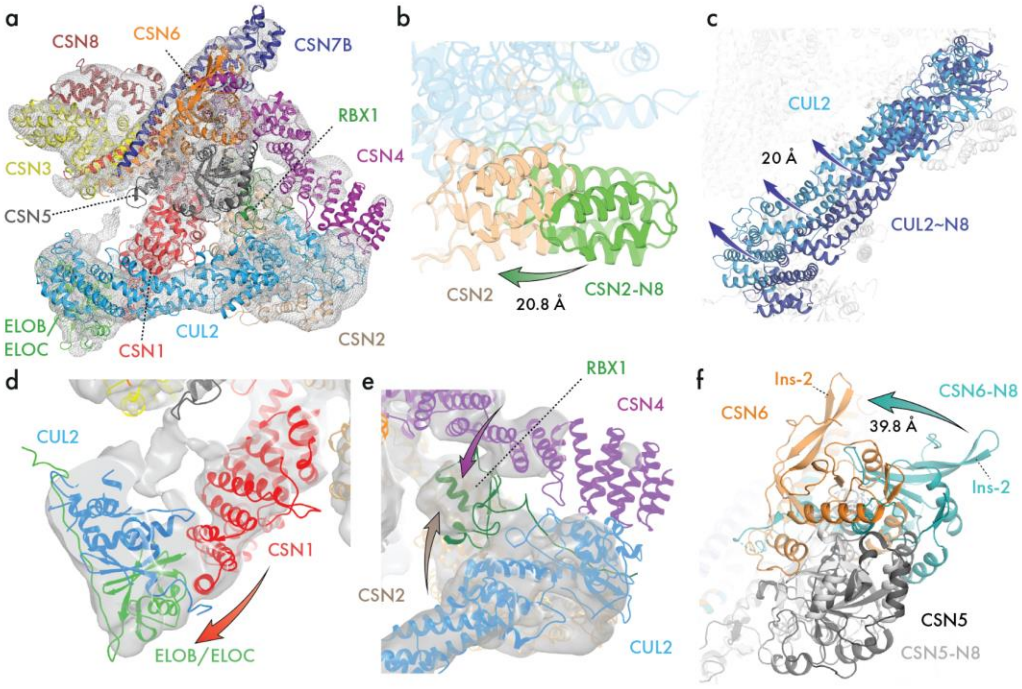


Figure 4.11. Structure of the deneddylated CSN–CRL2 complex. (a) The fitted density map of deneddylated CSN–CRL2 structure determined by combining cryo-EM and XL-MS. Alignment of the CSN C-terminal helical bundle from the neddylated and deneddylated holocomplexes reveals differences in (b) CSN2 and (c) CUL2. (d) CSN1-ELOB interface established from the rotation of CUL2 in (c). (e) RBX1 is clamped between CSN2 and CSN4. (f) Conformational changes in CSN6 in the absence of NEDD8.

The most striking conformational differences were observed in CSN6 (**Figure 4.11f**, **Supplementary Movie 6.1**). In the absence of NEDD8, CSN6 is dramatically shifted away from its position in the neddylated holocomplex by ~40 Å (**Figure 4.11f**). This previously unknown conformation of CSN6 differs from the conformation captured in our neddylated holocomplex, the CSN-CRL1~N8[222] and CSN-CRL4A~N8[223] structures (**Supplementary Figure 6.34-Supplementary Figure 6.35**). We compared the conformation of CSN6 seen in the structures of apo-CSN (4D10), CSN–CRL1~N8 (EMD-3401), CSN–CRL4A~N8 (EMD-3315), CSN–CRL4A^{DDB2}~N8 (EMD-3316) and our neddylated and non-neddylated CSN–CRL2 complexes through systematic structural alignments and measuring their pairwise root mean squared deviation (RMSD) (**Supplementary Figure 6.34-Supplementary Figure 6.35**). The conformation

of CSN6 in the non-neddylated CSN–CRL2 complex shows consistently high RMSD (16–25 Å) when compared with CSN6 in any of the other complexes indicating a high degree of conformational difference. CSN6 in the CSN–CRL2 structure has its Ins-2 loop dramatically shifted away from its apo-CSN conformation, pointing upwards to CSN7B (Supplementary Figure 6.34, Supplementary Movie 6.1). We anticipate that this conformation of CSN6 in our model of the non-neddylated CSN–CRL2 is attributed to the lack of NEDD8. Similar to the neddylated holocomplex, no significant changes were identified in CSN3, CSN7B and CSN8 subunits in the CSN–CRL2. Overall, comparison between the CSN–CRL2~N8 and CSN–CRL2 structure reveal significant conformational rearrangements in CSN5/CSN6 and the N-terminal domain of CSN2.

4.5.3 HDX-MS reveals a stepwise mechanism of CSN activation

Having determined the structures of the neddylated and deneddylated CSN–CRL2 complexes, we set off to characterise the local dynamics using HDX-MS. HDX-MS provides peptide-level information on the dynamics of proteins through monitoring the exchange events of amide hydrogens for bulk deuterium in the surrounding solution environment[65,142,182,245-248]. Here, we performed a set of two differential HDX-MS experiments to determine the effect of: (a) CRL2~N8 binding to CSN, denoted as Δ(CSN–CRL2~N8 - CSN), and (b) CRL2 binding to CSN, denoted as Δ(CSN^{WT}–CRL2 - CSN^{WT}) (Figure 4.12a, Supplementary Figure 6.36-Supplementary Figure 6.37). Regions that exhibit significant HDX differences brought about by the addition of the ligand (i.e. CRL2 and CRL2~N8) are labelled as stabilising (negative ΔHDX; coloured blue) or destabilising (positive ΔHDX; coloured red).

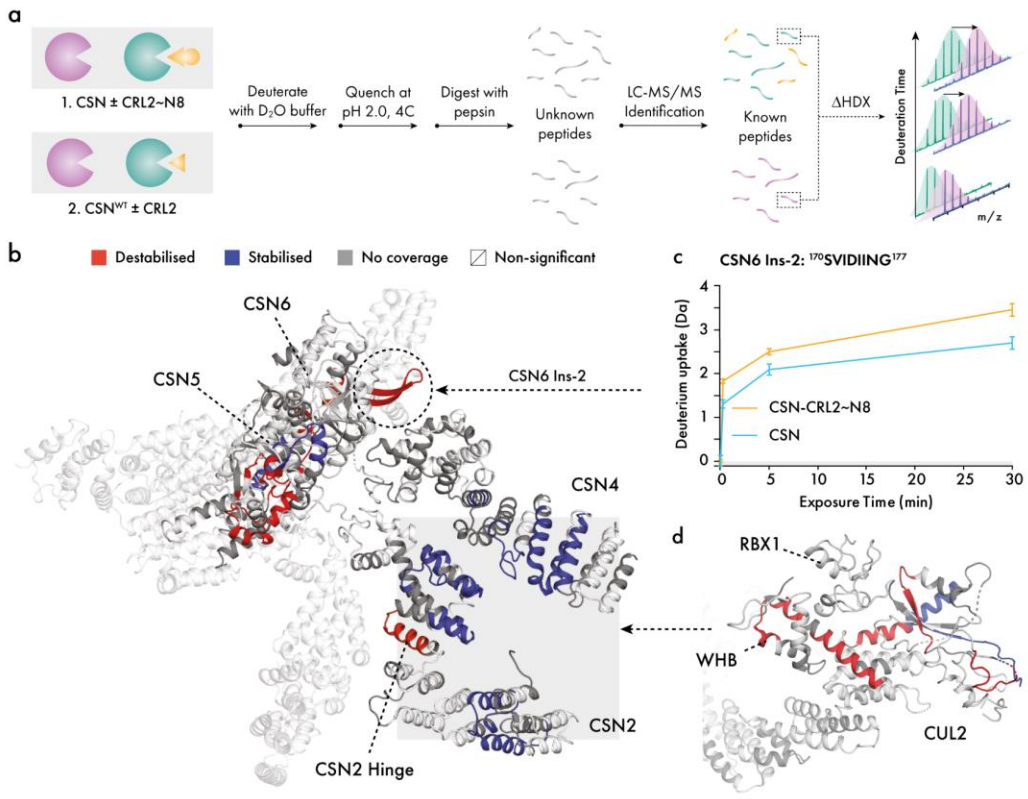


Figure 4.12. Effect of NEDD8 on the CSN4/CSN6 interface. (a) Investigating the effect of CRL2~N8 and CRL2 binding on CSN and CSN^{WT}, respectively. The two experiments involve deuterating the complexes for 0.25, 5 and 30 min time points, a quench step to halt the deuteration, and digestion to the peptide level. Peptides are then identified using liquid chromatography-tandem mass spectrometry (LC-MS/MS) and a database search. (b) Effect of CRL2~N8 binding on the CSN. (c) Relative deuterium uptake over 30 min for the CSN6 Ins-2 peptide (^{170}SVIDIING^{177}). Error bars indicate the deuterium uptake standard deviation for triplicate measurements (N = 3). (d) Effect of CSN on the C-terminal domain of CRL2~N8. Colour scheme follows that of (b).

In both Δ(CSN–CRL2~N8 - CSN) and Δ(CSN^{WT}–CRL2 - CSN^{WT}) experiments, extensive regions in the N-terminal helices of CSN2, CSN4 and the globular domain of RBX1 exhibited stabilisation upon the incubation of CSN with its CRL2~N8 and CRL2 substrates (Figure 4.12b, Supplementary Figure 6.38a-b). These observations are in line with the conformational clamping by CSN2/CSN4 onto the C-terminal of CRL2 as seen in the cryo-EM structures of neddylated and deneddylated complexes. Within CSN2 of both experiments, we observed significantly destabilised regions

around helical modules 6–9. These observations have likely identified the hinge points which permit the bending of CSN2 to clamp onto the CUL2 C-terminus (Figure 4.12b, Supplementary Figure 6.38a-b). Stabilised peptides belonging to CSN1 (143–156) and ELOB (17–25) were identified in the Δ(CSN^{WT}–CRL2 - CSN^{WT}) condition, indicating that an interface exists between CSN1–ELOB (Supplementary Figure 6.39). It is noted that since the CSN^{WT}–CRL2 sample used for HDX were not treated with GraFix (as they were for cryo-EM), it is unlikely that the CSN1–ELOB interface is caused by glutaraldehyde cross-linking. In the neddylated Δ(CSN–CRL2~N8 - CSN) condition, ELOB (17–25) was not found due to the lack of proteomic coverage. However, CSN1 (143–156) is stabilised, suggesting that the CSN1–ELOB interface is present in the neddylated holocomplex.

To assess the affinity between CSN4 and CRL2, we performed PLIMSTEX (protein–ligand interactions by mass spectrometry, titration and HDX) experiments (4.4 Materials and Methods). In these experiments, we performed HDX-MS of CSN or CSN^{WT} in the presence of increasing concentrations of CRL2 or CRL2~N8, to derive dissociation constants (Kd) between CSN4 and CRL2/CRL2~N8. PLIMSTEX differs from differential HDX-MS in several ways. First, no differential comparison is performed from PLIMSTEX. Second, PLIMSTEX experiments utilise a single deuteration time for all samples. PLIMSTEX requires a sufficiently long deuteration time to observe differences between holo and apo states, but short enough to prevent over-deuteration of interfaces as the complex naturally dissociates and re-associates. Finally, PLIMSTEX is a titration experiment which observes deuteration changes as a function of ligand concentration, while differential HDX-MS utilises a 1:1 molar ratio of our CSN and CRL2 substrates.

We tested the affinity of CSN4 for CRL2 or CRL2~N8 in CSN–CRL2~N8, CSN–CRL2 and CSN^{WT}–CRL2 complexes. Our PLIMSTEX experiments identified three regions of CSN4 which exhibit a dramatic decrease in deuterium uptake when exposed to the CRL2 substrate. These regions correspond to CSN4 α-helices which are in contact with CRL2~N8 in our cryo-EM fitted model of the CSN–CRL2~N8 (Supplementary Figure

6.40). The Kd measurements for the different regions of CSN4 in CSN^{WT}–CRL2 (11.3–34.5 nM), CSN–CRL2 (138.1–218.8 nM) and CSN–CRL2~N8 (118.9–389.0 nM) are each in the low nanomolar region, suggesting limited differences in local affinity brought about by changes such as the neddylation status of CRL2. These Kd values for the CSN–CRL2 system all fall within a similar overall range to the published global Kd to the CSN-CRL1 system[222] (1.6–310 nM; Supplementary Table 6.2), indicating a crucial role for CSN4 in stabilising CSN–CRL2. In addition, HDX-MS based Kd measurements also allowed us to localise individual regions of CSN4 responsible for interacting with CRL2 at the peptide level (Supplementary Figure 6.40).

4.5.4 Remodelling of the CSN5 active site in the presence of NEDD8

We next considered the release mechanism of the CSN5/CSN6 subunits of both the neddylated and deneddylated holocomplexes. In both Δ(CSN-CRL2~N8 – CSN) and Δ(CSN^{WT}-CRL2 – CSN^{WT}) experiments, the Ins-2 loop of CSN6 was destabilised, correlating with the release of CSN6 from its interface with CSN4 and in line with the allosteric activation mechanism of CSN by CRL4A[223] (Figure 4.13a-b, i). An interesting difference between the neddylated and deneddylated complexes is that the CSN6 α4 helix is destabilised only in the absence of NEDD8 (Figure 4.13b, i). Similarly, the CSN5 α7 helix is also destabilised in both neddylated and deneddylated conditions (Figure 4.13b, ii). The CSN6 α4 and CSN5 α7 helices are topologically knotted in the CSN5/CSN6 heterodimer and tether the globular domains of CSN5/CSN6 to the C-terminal helical bundle[217]. These observations suggest that structural changes are required in the helical knot to bring about release of the CSN5/CSN6 globular domains from their apo conformation.

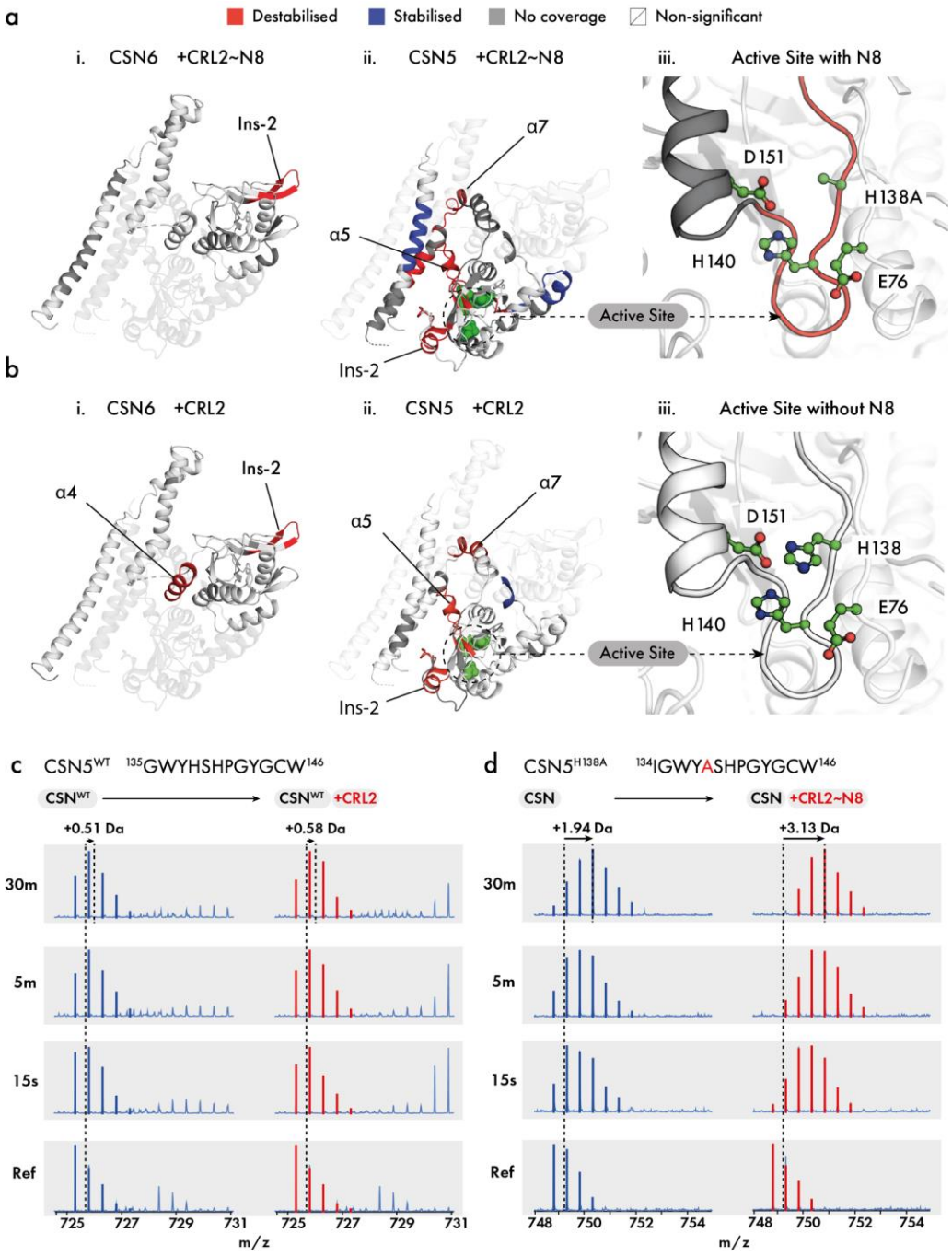


Figure 4.13. Conformational response of CSN5/CSN6 to NEDD8. Differential HDX-MS of (a) Δ(CSN–CRL2~N8 - CSN) and (b) Δ(CSN^{WT}–CRL2 - CSN^{WT}). Regions exhibiting significant deuterium uptake differences in (i) CSN6, (ii) CSN5 and (iii) CSN5 active site are highlighted in red for destabilised and blue for stabilised areas. Regions without coverage or exhibit non-significant changes are in grey and white, respectively. (c–d) Deuteration profiles of CSN5 active site (as shown in a–b iii). Profiles of the c non-neddylated and d neddylated holocomplex (red) are compared with deuteration profiles of the isolated CSN (blue). The ion distribution of active site peptide is shown for (c) and (d) across reference 0 s, 15 s, 5 m and 30 m timepoints. The mass of the non-deuterated reference peptide is shown by the dotted line. A second dotted line for the holocomplexes (red)

indicates the mass of the peptide at the 30 m timepoint. The relative deuterium changes of the active site peptide in (c) following binding of CRL2 is negligible, while in (d), presence of CRL2~N8 leads to a significant increase in mass.

Another finding is that we identified destabilisation in the Ins-2 loop of CSN5 (Figure 4.13a-b, i). The Ins-2 loop of CSN5 has a lesser understood role in CSN activation. In isolated CSN5, the Ins-2 loop is highly disordered[225] (Supplementary Figure 6.41a), while it folds into a helical-loop structure when incorporated into the CSN[217] (Supplementary Figure 6.41b). Accompanying the changes in the CSN5 Ins-2 loop, in both comparative HDX-MS experiments, we detected destabilisation of the α5 helix area which surrounds the CSN5 active site (Figure 4.13a-b, ii). The changes in both the CSN5 Ins-2 and α5 helix indicate a major conformational remodelling in the area adjacent to the CSN5 active site, which can be triggered through the binding of both CRL2 or CRL2~N8 to the CSN in a NEDD8-independent manner. It is only in the presence of NEDD8, that the CSN5 active site is further destabilised suggesting that in a final activation step, NEDD8 induces conformational changes in the active site itself (Figure 4.13). To eliminate the possibility that the observed changes in the CSN5 active site are due to the H138A point mutation, we compared the deuterium uptake profiles of CSN5 from apo-CSN^{WT} and CSN^{5H138A} constructs (Supplementary Figure 6.42). Calculating the deuterium uptake differences between peptides of the CSN5^{WT} and CSN^{5H138A} and visualising this through a Woods plot, identified no significant uptake differences (Supplementary Figure 6.42). With this considered, the deprotection observed in the CSN5 active site of the CSN–CRL2~N8 complex can be seen to result from the binding of CRL2~N8 and not the H138A mutation of the CSN5 active site.

4.6 Discussion

Here we have combined EM and MS analyses to provide insights into the mediation of CRL2 by the CSN. We have described the molecular structures of CSN–CRL2~N8 and its deneddylated CSN–CRL2 counterpart. Furthermore, we combined cryo-EM

maps with comparative HDX-MS to expand on the stepwise activation mechanism of the CSN, involving a conformational network of both NEDD8-independent and dependent stages. We suggest that the steps which lead to deneddylation are mostly NEDD8-independent, except for the remodelling of the CSN5 active site which requires NEDD8 to encounter the CSN5 active site.

Our map of the deneddylated CSN–CRL2 holocomplex represents a complex in which the CSN is still associated with its CRL2 reaction product. Resolving this structure has provided several important details into how activation of the CSN is achieved. Our comparison of the neddylated and deneddylated holocomplex structures indicated that the CSN2 contacts the CRL2 C-terminal domain in a slightly different conformation to when the CRL2 is modified with NEDD8. Between both neddylated and deneddylated conformations, we suggest that the clamping by CSN2 involves destabilization of the CSN2 helical modules 6–9 which function possibly as a hinge that allows the CSN2 to bend upwards towards the CRL2. The plasticity of these N-terminal helices presumably permits the binding of deneddylated and alternative Cullin isoforms. HDX-MS further indicates that RBX1 and CSN4 form an interface, which is more prominent in the absence of NEDD8, as shown through stabilisation of the two interfaces (Supplementary Figure 6.38b). Overall, the conformational variations seen in the CSN2 N-terminal helical modules (Figure 4.11b), the bend of CRL2 (Figure 4.11c) and the HDX differences in CSN4/RBX1 (Supplementary Figure 6.38b) may therefore be attributable to the seemingly promiscuous affinity that allows CSN2 to bind to each of the different Cullins regardless of neddylation.

We uncovered structural and dynamical aspects of both the neddylated and deneddylated CSN–CRL2 complexes. Beginning our interpretation from valuable studies of the CSN–CRL1~N8 and CSN–CRL4A~N8 systems, we propose several major conformational switches of the CSN which must be activated by the CRL2 substrate to bring about deneddylation. While some of these steps are conserved ubiquitously among other CSN-CRL complexes (e.g. CSN2/CSN4 clamping in CSN–

CRL1~N8 and CSN–CRL4A~N8), our study suggests additional steps for the CRL2-bound CSN complex (Figure 4.14). In the first activation step, the CSN and CRL2~N8 associate through major conformational changes in CSN2 and CSN4, which clamp onto the CRL2 (Figure 4.14a-b). Our data suggest that the conformational change in CSN4 breaks its interface with CSN6 through the CSN6 Ins-2 loop and with the eventual release of the CSN5/CSN6 heterodimer (Figure 4.14c). Removal of the CSN6 Ins-2 loop has been shown in CSN-CRL1 to disrupt the CSN4-CSN6 interface, leading to sustained enzymatic activity of the CSN[217]. In future studies, targeted deletion of the CSN6 Ins-2 loop can be performed alongside differential HDX-MS for the CSN-CRL2 system, to further probe the differences in active site remodelling of CSN5 as a result of disrupting the CSN4-CSN6 interface.

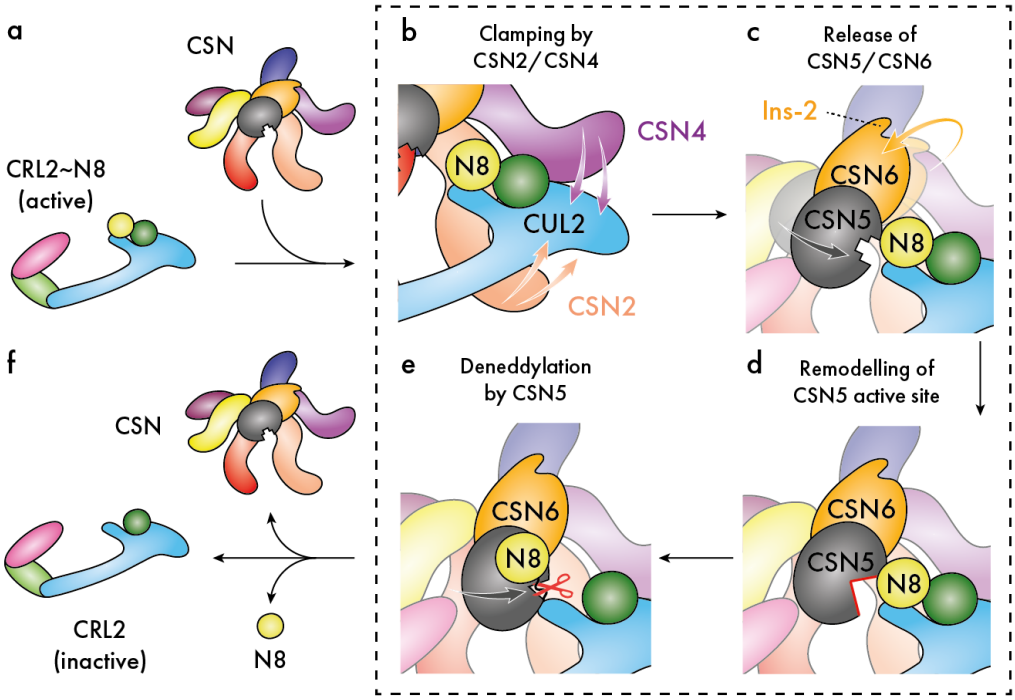


Figure 4.14. Schematic of CRL2 regulation by the CSN. (a) Neddylated (active) CRL2~N8 is regulated by the CSN. (b) CRL2~N8 is bound by the CSN, principally through clamping by the extended N-terminal helical modules of subunits CSN2 and CSN4. (c) The topological changes that occur in CSN4 upon CRL2 binding result in detachment of the CSN6 Ins-2 loop and subsequent release of the CSN5/CSN6 heterodimer in a NEDD8-independent step. (d) The presence of NEDD8 triggers remodelling of the CSN5 active site, (e) permitting deneddylation of CRL2. (f) Cleavage of NEDD8 results in dissociation of the CSN and inactive CRL2.

The release of CSN5/CSN6 appears consistent with the destabilised knotted helices of CSN6 that our HDX has identified. It is plausible that these two helices function as the mechanical hinges which allow the CSN5/CSN6 to be released from their auto-inhibited conformations but remain tethered to the rest of the CSN. Although the resolution presented by CSN5 in our cryo-EM structures prevents us from making molecular level observations, our HDX data can provide local detail for the CSN5 active site. The release of CSN5/CSN6 is accompanied by HDX changes in areas surrounding the CSN5 active site, including the CSN5 Ins-2 loop. Up to here, the changes experienced by the CSN can be brought about in a NEDD8-independent manner. In the next stage, the presence of NEDD8 acts as a selectivity filter which results in remodelling of the CSN5 active site itself (Figure 4.14d). These changes presumably expose the CSN5 JAMM ligands of the metalloprotease site and allow subsequent deneddylation to occur (Figure 4.14e). Finally, deneddylation ensues with the cleavage of NEDD8 from CRL2 and the dissociation of the complex (Figure 4.14f). The fact that CSN can then re-associate with its CRL2 reaction product following dissociation, as shown by our study and structure of the CSN–CRL2, suggests that the non-neddylated complex may possess an alternative role to deneddylation. By associating with non-neddylated CRLs, the CSN sterically blocks access of both ubiquitination E2 enzymes and substrates to the CRLs[219,223]. Further studies will be required to fully understand the deneddylation-independent roles of the CSN. Interestingly, a comparison of our non-neddylated CSN–CRL2 with the apo-CSN and neddylated CSN–CRL4~N8 and CSN–CRL2~N8 complexes, highlighted the dramatic differences in the conformation of CSN6. In the absence of NEDD8, the CSN5/CSN6 are released much further than any other observed conformation of CSN6. We hypothesise that this difference may arise from the lack of steric hinderance usually presented by NEDD8, allowing CSN5/CSN6 to rotate and meet the Cullin scaffold much closer than in the neddylated structure.

In our study we have made interpretations of the CSN2 and CSN4 conformational clamping through comparing our cryo-EM holocomplexes with crystal structures of the apo-CSN. From reviewing the conformations of all nine currently available independent copies of the apo-CSN molecule it is apparent that CSN2 and CSN4 appear to exist in a range of open conformations that are quite distinct from the closed conformations that we observed in CSN–CRL2 complexes (Supplementary Figure 6.25), and the similarly closed conformations observed in CSN complexes with CRL1~N8 and CRL4A~N8. However, each of the apo-CSN crystal structures is characterised by crystal contacts involving CSN2 and CSN4 which in principle could have biased their conformations.

Therefore, we used XL-MS to further monitor the conformation of CSN4 within the solution structure of apo-CSN and found that both open and closed conformations of apo-CSN are required to satisfy a number of unique cross-links (Supplementary Figure 6.35). Since our study used lysine cross-links, our distance measurements do not allow for a high level of scrutiny in our modelling due to the uncertainty of lysine side chain rotamers in our models. Nonetheless, given that the violated cross-link distances are well above 40–50 Å for one model and close to 35 Å for the other (Supplementary Figure 6.31), we can assume that both open and closed conformations are simultaneously represented within the cross-link ensemble. Thus it seems likely that crystal formation has stabilised CSN conformations already present in solution. Hence, while we cannot completely rule out the possibility that the crystallographically defined open conformations of apo-CSN are atypical, it seems likely that the solution structure of the apo-CSN corresponds to an ensemble containing a range of open structures including those seen crystallographically as well as some closed or nearly closed structures. In comparison, CSN2 and CSN4 in the cryo-EM structures of CSN–CRL2 complexes adopt consistently closed conformations sandwiching the C-terminal end of CUL2 with rather little apparent conformational variation. It is likely here that conformational variation is limited by interactions formed with CUL2. Thus it would appear that the interaction between

CSN and CRL2 involves the transition of CSN from an ensemble of conformers with substantial variation in the distance between CSN2 and CSN4 to a bound complex in which CSN2 and CSN4 clamp onto the C-terminal region of CUL2 which is characterised by greatly reduced conformational variation.

To further explore the nature of such clamped complexes, we additionally performed PLIMSTEX experiments which provided localised affinity values between CSN4 and CUL2 for each combination of the CSN–CRL2. From a functional perspective, our PLIMSTEX experiments also determined that the H138A mutation of CSN5 leads to measurable changes in Kd between CSN4 and CRL2. It is also important to note that the Kd values presented are not representative of the global Kd between CSN and CRL2 complexes but are only local affinities between the CSN4 and CRL2 subunits. PLIMSTEX experiments require that the labelling time of the experiment is carefully selected based on preliminary experiments[249]. While powerful, a concern is that if the labelling time is too long, protein interfaces may become over-deuterated, leading to an underestimated deuterium uptake decrease and a higher Kd value. Likewise, some interfaces may appear invisible if the labelling time is not sufficiently long to allow deuteration to occur.

In both neddylated and non-neddylated CSN-CRL2 complexes, we identified an interface between CSN1 and ELOB. So far, CSN1 has been shown to interact with Skp2 and Fbw7 in CRL1[219] and DDB1 in CRL4A holocomplexes[223]. Mutations of the CSN1-DDB1 interface did not affect binding to the CSN nor perturb deneddylation activity[223]. In the absence of Cullin-1 or RBX1, CRL1 substrate receptors do not associate with the CSN1[219]. Although no CSN activator role has been assigned to CSN1-substrate receptor interactions, their presence increases the interface between CSN and their CRL substrates[219] and may stabilise these interactions.

Furthermore, there may be additional roles for the CSN as suggested by the compositional plasticity seen in our CSN–CRL2~N8 classes. In our map of the CSN–CRL2~N8, the Cullin arm was observed to shift downwards away from CSN3 in maps

deficient of the VHL substrate receptor (Supplementary Figure 6.43), consistent with the coupling of VHL binding to a conformational change in the rest of the CRL2 portion of the protein. This type of behavior may allow the CRL2 to adapt to individual substrates and substrate receptors which vary in size and geometry as has been suggested with the CRL4A system[223]. In addition, it may reflect changes associated with remodeling of the CRL2 by the dissociation of the VHL and the binding of alternative substrate receptors. In future work we will seek to determine whether the CSN can mediate substrate receptor dissociation.

Overall, our study has provided greater detail into the role of CRL2 and NEDD8 in regulating the activation mechanism of CSN. We propose that the series of mechanistic responses of the CSN that lead up to deneddylation, can be triggered even by the CRL2 reaction product in a NEDD8-independent manner. The presence of NEDD8 on the activated CRL2 substrate would then trigger the remodeling within the catalytic site of the CSN5 subunit during the final stage of CSN activation. We envision that this type of mechanism would have implications for the entire family of CRL proteins and their regulatory relationship with the CSN. Our study therefore provides a template not only for assisting investigations of other CRL-based systems but also for bringing together data from different structural biology techniques that otherwise will be reported independently.

Chapter 5: Conclusions and Future Outlook

From the perspective of a structural biologist, the discovery of a protein's structure can be thought of as overcoming a major energy barrier in the understanding of its function. The challenge of protein structure determination in the past has been served by techniques such as X-ray crystallography and NMR. The advent of cryo-EM and its "resolution revolution" has brought with it a new era of structural determination, particularly allowing access to previously insuperable systems such as dynamic macromolecules and membrane protein systems. MS has served a complementary role to these techniques by providing a toolkit that can probe multiple structural and dynamic dimensions of biological molecules. Combined interpretation of atomistic structures with MS data provides a more complete understanding than reviewing either of the techniques alone and can be streamlined through the use of computational strategies as demonstrated in this thesis. As the structures of increasing numbers of proteins become available, the understanding of protein function will be limited by the next energy barrier, in the form of its conformational landscape and interactome. Mapping out the dynamic mechanisms and interactions of proteins will lead to better knowledge of cellular function and its relationship to health and disease.

In this thesis, I have demonstrated three methods of dynamic characterization of proteins. In the first chapter, a computational workflow capable of replicating the gas phase structures of large flexible proteins was demonstrated. This methodology combined modelling strategies with MD simulations and produced representative models of each of the IgG1-4 isotypes based on agreement with IM-MS measurements. Previous attempts to replicate the gas phase topologies of IgG1 directly from their crystal structures were less successful. The utility of the methodology is two-fold. First, it highlights the need to consider the solution

ensemble of flexible molecules prior to any gas phase simulations. In solution, flexible proteins occupy a wide continuum of conformations. We propose a model envisioning that flexible proteins are coaxed into occupying a subset of compact solution conformations during ESI droplet shrinkage, and that these compact conformations later populate the gas phase distribution. Second, we demonstrate that a combination of a topological sampling with short gas phase MD (~5 ns) is a computationally inexpensive alternative to trajectory stitching methods. This approach can be in theory applied to any proteins which derive their flexibility from hinge or linker regions.

Since structural collapse in the gas phase is directly related to flexibility in solution, this methodology opens exciting avenues into the use of IM-MS for characterisation of protein flexibility. One such application may be in the development of biotherapeutics whereby our method can be used in a differential manner to characterise changes in the conformational space of antibodies as a function of mutation or drug binding (Figure 5.1). Such integration may streamline the drug screening process by providing existing IM-MS workflows with the means of generating representative structures for detailed comparisons.

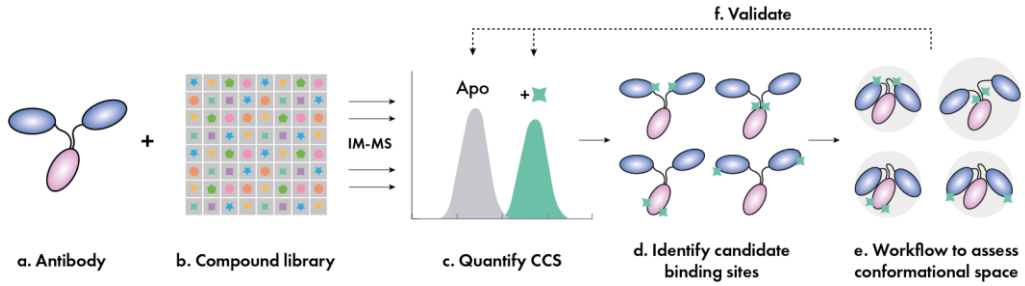


Figure 5.1. Modelling workflow can be used to assess differences in conformational space of antibodies upon drug binding. (a-b) Antibodies and drug compounds can be screened using IM-MS. (c) Identify compounds leading to changes in conformation. (d) Identify candidate binding sites, e.g. through docking analysis. (e) Application of our computational workflow to assess gas phase conformations of antibody-drug pairs. (f) Validate possible candidate sites through matching *CCS*.

The use of HDX-MS for protein characterisation has gained traction over the recent years due to its ability to capture dynamic aspects of protein behaviour. Developments into instrumentation capable of performing high-throughput HDX-MS, has been paralleled with studies of increasing biological complexity, and equal developments into data analysis software. The timeline of HDX-MS over the last few decades has seen multiple shifts in the technique's bottleneck, from first, data acquisition speeds, to data processing and now finally to meaningful visualisation and interpretation. Chapter 3 describes the development of Deuteros – statistical analysis and visualisation software designed to facilitate interpretation of HDX-MS results. Analysis of protein behaviour in differential comparisons remains one of the most commonly applied methods of HDX-MS. The extraction of potentially interesting peptides using statistical methods, is one of the strengths of Deuteros through its Woods and volcano plotting facilities. While the software still has much room for improvement, its utility is evident in the growing number of publications that showcase Deuteros-analysed data. Improvements to the backend and analysis workflow such as those documented in this thesis, will hopefully continue to be of use to the HDX-MS community upon the release of Deuteros 2.0. We anticipate that the new features of Deuteros 2.0 will further enhance interpretation of HDX-MS data leading to fruitful discoveries in protein functionality.

Dynamic behaviours occur not only in flexible molecules such as antibodies, but also manifest during transitions along a protein's mechanistic landscape. An example of a complex reaction pathway involving several conformational transitions, can be found in the CSN complex of the UPS. In Chapter 4, we demonstrated the utility of a combinatory structural MS and cryo-EM approach to determine the precise steps of the deneddylation cascade of the CSN. The activation of the CSN involves a dynamic network of conformational events, relayed through binding to neddylated CRLs. Uniting several structural techniques offers a synergistic advantage since the shortcomings of individual techniques can be compensated for by information contained in others. Several techniques offering redundancy to observations also

gives confidence to conclusions and can avoid false positives from being reported. Cryo-EM provided unprecedented resolution into the topology and arrangement of the CSN, however was unable to resolve dynamic regions critical to the understanding of how deneddylation occurs. We utilised XL-MS to capture proximity information on the dynamic regions and used the resultant cross-links in computational modelling to guide their placement. Sub-complexes of the CSN were identified in both cryo-EM and native MS spectra, suggesting their natural occurrence within the distribution of CSN structures in solution. Further differential HDX-MS provided a method of capturing dynamic events upon binding of both the neddylated and non-neddylated CRL2 to CSN, and biologically interesting regions were extracted using Deuteros. Overall, our multi-technique approach has determined several novel and key events that occur during activation of the CSN. Our study supports the notion of conserved interactions between the CSN and its cognate CRLs, but also demonstrates the utility and often, necessity of complex characterisation from multiple structural angles.

In summary, this thesis has contributed towards developing methodologies for the characterisation of protein dynamics, through the integration of structural MS and computational modelling. As structural biology progresses into the “dynamic era”, methods capable of dynamic characterisation will be indispensable to understanding the mechanisms of increasingly complex biological systems and even interactions at the cellular level. The potential applications of integrative modelling are endless, and will no doubt stimulate further discussions, spark collaborations and lead to breakthroughs in biology.

Bibliography

1 Wortmann, A. *et al.* Shrinking droplets in electrospray ionization and their influence on chemical equilibria. *J Am Soc Mass Spectrom* **18**, 385-393, doi:10.1016/j.jasms.2006.10.010 (2007).

2 Wilm, M. Principles of electrospray ionization. *Mol Cell Proteomics* **10**, M111 009407, doi:10.1074/mcp.M111.009407 (2011).

3 Hogan, C. J., Jr., Carroll, J. A., Rohrs, H. W., Biswas, P. & Gross, M. L. Combined charged residue-field emission model of macromolecular electrospray ionization. *Anal Chem* **81**, 369-377, doi:10.1021/ac8016532 (2009).

4 Konermann, L., Ahadi, E., Rodriguez, A. D. & Vahidi, S. Unraveling the mechanism of electrospray ionization. *Anal Chem* **85**, 2-9, doi:10.1021/ac302789c (2013).

5 Dole, M. *et al.* Molecular Beams of Macroions. *The Journal of Chemical Physics* **49**, 2240-2249, doi:10.1063/1.1670391 (1968).

6 Grandori, R. Origin of the conformation dependence of protein charge-state distributions in electrospray ionization mass spectrometry. *J Mass Spectrom* **38**, 11-15, doi:10.1002/jms.390 (2003).

7 Patriksson, A., Marklund, E. & van der Spoel, D. Protein structures under electrospray conditions. *Biochemistry* **46**, 933-945, doi:10.1021/bi061182y (2007).

8 Leney, A. C. & Heck, A. J. Native Mass Spectrometry: What is in the Name? *J Am Soc Mass Spectrom* **28**, 5-13, doi:10.1007/s13361-016-1545-3 (2017).

9 Ahdash, Z. *et al.* Mechanistic insight into the assembly of the HerA-NurA helicase-nuclease DNA end resection complex. *Nucleic Acids Res* **45**, 12025-12038, doi:10.1093/nar/gkx890 (2017).

10 Heck, A. J. & Van Den Heuvel, R. H. Investigation of intact protein complexes by mass spectrometry. *Mass Spectrom Rev* **23**, 368-389, doi:10.1002/mas.10081 (2004).

11 Pyle, E. *et al.* Structural Lipids Enable the Formation of Functional Oligomers of the Eukaryotic Purine Symporter UapA. *Cell Chem Biol* **25**, 840-848 e844, doi:10.1016/j.chembiol.2018.03.011 (2018).

12 van den Heuvel, R. H. & Heck, A. J. Native protein mass spectrometry: from intact oligomers to functional machineries. *Curr Opin Chem Biol* **8**, 519-526, doi:10.1016/j.cbpa.2004.08.006 (2004).

13 Hall, Z. & Robinson, C. V. Do charge state signatures guarantee protein conformations? *J Am Soc Mass Spectrom* **23**, 1161-1168, doi:10.1007/s13361-012-0393-z (2012).

14 Gabelica, V. & Marklund, E. Fundamentals of ion mobility spectrometry. *Curr Opin Chem Biol* **42**, 51-59, doi:10.1016/j.cbpa.2017.10.022 (2018).

15 Ruotolo, B. T., Benesch, J. L., Sandercock, A. M., Hyung, S. J. & Robinson, C. V. Ion mobility-mass spectrometry analysis of large protein complexes. *Nat Protoc* **3**, 1139-1152, doi:10.1038/nprot.2008.78 (2008).

16 Zhou, Z., Shen, X., Tu, J. & Zhu, Z. J. Large-Scale Prediction of Collision Cross-Section Values for Metabolites in Ion Mobility-Mass Spectrometry. *Anal Chem* **88**, 11084-11091, doi:10.1021/acs.analchem.6b03091 (2016).

17 Gabelica, V. *et al.* Recommendations for reporting ion mobility Mass Spectrometry measurements. *Mass Spectrom Rev* **38**, 291-320, doi:10.1002/mas.21585 (2019).

18 Allison, T. M. *et al.* Quantifying the stabilizing effects of protein-ligand interactions in the gas phase. *Nat Commun* **6**, 8551, doi:10.1038/ncomms9551 (2015).
19 Mack, E. Average cross-sectional areas of molecules by gaseous diffusion methods. *Journal of the American Chemical Society* **47**, 2468-2482, doi:DOI 10.1021/ja01687a007 (1925).
20 Mesleh, M. F., Hunter, J. M., Shvartsburg, A. A., Schatz, G. C. & Jarrold, M. F. Structural information from ion mobility measurements: Effects of the long-range potential. *J Phys Chem-Us* **100**, 16082-16086, doi:DOI 10.1021/jp961623v (1996).
21 Jurneczko, E. & Barran, P. E. How useful is ion mobility mass spectrometry for structural biology? The relationship between protein crystal structures and their collision cross sections in the gas phase. *Analyst* **136**, 20-28, doi:10.1039/c0an00373e (2011).
22 Benesch, J. L. & Ruotolo, B. T. Mass spectrometry: come of age for structural and dynamical biology. *Curr Opin Struct Biol* **21**, 641-649, doi:10.1016/j.sbi.2011.08.002 (2011).
23 Marklund, E. G., Degiacomi, M. T., Robinson, C. V., Baldwin, A. J. & Benesch, J. L. Collision cross sections for structural proteomics. *Structure* **23**, 791-799, doi:10.1016/j.str.2015.02.010 (2015).
24 Bleiholder, C., Wyttenbach, T. & Bowers, M. T. A novel projection approximation algorithm for the fast and accurate computation of molecular collision cross sections (I). Method. *Int J Mass Spectrom* **308**, 1-10, doi:10.1016/j.ijms.2011.06.014 (2011).
25 Lee, J. W., Lee, H. H. L., Davidson, K. L., Bush, M. F. & Kim, H. I. Structural characterization of small molecular ions by ion mobility mass spectrometry in nitrogen drift gas: improving the accuracy of trajectory method calculations. *Analyst* **143**, 1786-1796, doi:10.1039/C8AN00270C (2018).
26 Shvartsburg, A. A. & Jarrold, M. F. An exact hard-spheres scattering model for the mobilities of polyatomic ions. *Chem Phys Lett* **261**, 86-91, doi:Doi 10.1016/0009-2614(96)00941-4 (1996).
27 Ewing, S. A., Donor, M. T., Wilson, J. W. & Prell, J. S. Collidoscope: An Improved Tool for Computing Collisional Cross-Sections with the Trajectory Method. *J Am Soc Mass Spectrom* **28**, 587-596, doi:10.1007/s13361-017-1594-2 (2017).
28 Devine, P. W. A. *et al.* Investigating the Structural Compaction of Biomolecules Upon Transition to the Gas-Phase Using ESI-TWIMS-MS. *J Am Soc Mass Spectrom* **28**, 1855-1862, doi:10.1007/s13361-017-1689-9 (2017).
29 Pacholarz, K. J. *et al.* Dynamics of intact immunoglobulin G explored by drift-tube ion-mobility mass spectrometry and molecular modeling. *Angew Chem Int Ed Engl* **53**, 7765-7769, doi:10.1002/anie.201402863 (2014).
30 Scott, D., Layfield, R. & Oldham, N. J. Ion mobility-mass spectrometry reveals conformational flexibility in the deubiquitinating enzyme USP5. *Proteomics* **15**, 2835-2841, doi:10.1002/pmic.201400457 (2015).
31 Brockwell, D. J. *et al.* The effect of core destabilization on the mechanical resistance of I27. *Biophys J* **83**, 458-472, doi:10.1016/S0006-3495(02)75182-5 (2002).
32 Breuker, K. & McLafferty, F. W. Stepwise evolution of protein native structure with electrospray into the gas phase, 10(-12) to 10(2) s. *Proc Natl Acad Sci U S A* **105**, 18145-18152, doi:10.1073/pnas.0807005105 (2008).
33 Koeniger, S. L., Merenbloom, S. I. & Clemmer, D. E. Evidence for many resolvable structures within conformation types of electrosprayed ubiquitin ions. *J Phys Chem B* **110**, 7017-7021, doi:10.1021/jp056165h (2006).
34 Karplus, M. & Petsko, G. A. Molecular dynamics simulations in biology. *Nature* **347**, 631-639, doi:10.1038/347631a0 (1990).

35 Sultan, M. M., Denny, R. A., Unwalla, R., Lovering, F. & Pande, V. S. Millisecond dynamics of BTK reveal kinome-wide conformational plasticity within the apo kinase domain. *Sci Rep* **7**, 15604, doi:10.1038/s41598-017-10697-0 (2017).
36 Brooks, B. R. *et al.* Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *Journal of Computational Chemistry* **4**, 187-217, doi:DOI 10.1002/jcc.540040211 (1983).
37 Guvench, O. & MacKerell, A. D. in *Molecular Modeling of Proteins* (ed Andreas Kukol) 63-88 (Humana Press, 2008).
38 Lopes, P. E., Guvench, O. & MacKerell, A. D., Jr. Current status of protein force fields for molecular dynamics simulations. *Methods Mol Biol* **1215**, 47-71, doi:10.1007/978-1-4939-1465-4_3 (2015).
39 Nutt, D. R. & Smith, J. C. Molecular Dynamics Simulations of Proteins: Can the Explicit Water Model Be Varied? *J Chem Theory Comput* **3**, 1550-1560, doi:10.1021/ct700053u (2007).
40 Marklund, E. G. & Benesch, J. L. Weighing-up protein dynamics: the combination of native mass spectrometry and molecular dynamics simulations. *Curr Opin Struct Biol* **54**, 50-58, doi:10.1016/j.sbi.2018.12.011 (2019).
41 Konermann, L., Metwally, H., McAllister, R. G. & Popa, V. How to run molecular dynamics simulations on electrospray droplets and gas phase proteins: Basic guidelines and selected applications. *Methods* **144**, 104-112, doi:10.1016/j.ymeth.2018.04.010 (2018).
42 McAllister, R. G., Metwally, H., Sun, Y. & Konermann, L. Release of Native-like Gaseous Proteins from Electrospray Droplets via the Charged Residue Mechanism: Insights from Molecular Dynamics Simulations. *J Am Chem Soc* **137**, 12667-12676, doi:10.1021/jacs.5b07913 (2015).
43 Meyer, T., Gabelica, V., Grubmüller, H. & Orozco, M. Proteins in the gas phase. *Wiley Interdisciplinary Reviews: Computational Molecular Science* **3**, 408-425, doi:10.1002/wcms.1130 (2013).
44 Popa, V., Trecroce, D. A., McAllister, R. G. & Konermann, L. Collision-Induced Dissociation of Electrosprayed Protein Complexes: An All-Atom Molecular Dynamics Model with Mobile Protons. *J Phys Chem B* **120**, 5114-5124, doi:10.1021/acs.jpcb.6b03035 (2016).
45 Sinz, A. Cross-Linking/Mass Spectrometry for Studying Protein Structures and Protein-Protein Interactions: Where Are We Now and Where Should We Go from Here? *Angew Chem Int Ed Engl* **57**, 6390-6396, doi:10.1002/anie.201709559 (2018).
46 Peng, T., Yuan, X. & Hang, H. C. Turning the spotlight on protein-lipid interactions in cells. *Curr Opin Chem Biol* **21**, 144-153, doi:10.1016/j.cbpa.2014.07.015 (2014).
47 Tretyakova, N. Y., Groehler, A. t. & Ji, S. DNA-Protein Cross-Links: Formation, Structural Identities, and Biological Outcomes. *Acc Chem Res* **48**, 1631-1644, doi:10.1021/acs.accounts.5b00056 (2015).
48 Madler, S., Bich, C., Touboul, D. & Zenobi, R. Chemical cross-linking with NHS esters: a systematic study on amino acid reactivities. *J Mass Spectrom* **44**, 694-706, doi:10.1002/jms.1544 (2009).
49 Preston, G. W. & Wilson, A. J. Photo-induced covalent cross-linking for the analysis of biomolecular interactions. *Chem Soc Rev* **42**, 3289-3301, doi:10.1039/c3cs35459h (2013).
50 Yang, T., Li, X. M., Bao, X., Fung, Y. M. & Li, X. D. Photo-lysine captures proteins that bind lysine post-translational modifications. *Nat Chem Biol* **12**, 70-72, doi:10.1038/nchembio.1990 (2016).
51 Olsen, J. V., Ong, S. E. & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol Cell Proteomics* **3**, 608-614, doi:10.1074/mcp.T400003-MCP200 (2004).
52 Holding, A. N. XL-MS: Protein cross-linking coupled with mass spectrometry. *Methods* **89**, 54-63, doi:10.1016/j.ymeth.2015.06.010 (2015).

53 Leitner, A. *et al.* Expanding the chemical cross-linking toolbox by the use of multiple proteases and enrichment by size exclusion chromatography. *Mol Cell Proteomics* **11**, M111 014126, doi:10.1074/mcp.M111.014126 (2012).
54 Choksawangkarn, W., Edwards, N., Wang, Y., Gutierrez, P. & Fenselau, C. Comparative study of workflows optimized for in-gel, in-solution, and on-filter proteolysis in the analysis of plasma membrane proteins. *J Proteome Res* **11**, 3030-3034, doi:10.1021/pr300188b (2012).
55 Leitner, A. Cross-linking and other structural proteomics techniques: how chemistry is enabling mass spectrometry applications in structural biology. *Chem Sci* **7**, 4792-4803, doi:10.1039/c5sc04196a (2016).
56 Schilling, B., Row, R. H., Gibson, B. W., Guo, X. & Young, M. M. MS2Assign, automated assignment and nomenclature of tandem mass spectra of chemically crosslinked peptides. *Journal of the American Society for Mass Spectrometry* **14**, 834-850, doi:10.1016/s1044-0305(03)00327-1 (2003).
57 Rappsilber, J. The beginning of a beautiful friendship: cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *J Struct Biol* **173**, 530-540, doi:10.1016/j.jsb.2010.10.014 (2011).
58 Brewis, I. A. & Brennan, P. Proteomics technologies for the global identification and quantification of proteins. *Adv Protein Chem Struct Biol* **80**, 1-44, doi:10.1016/B978-0-12-381264-3.00001-1 (2010).
59 Quan, L. D. L., M. CID,ETD and HCD Fragmentation to Study Protein Post-Translational Modifications. *Modern Chemistry & Applications* **01**, doi:10.4172/2329-6798.1000e102 (2013).
60 Chen, Z. L. *et al.* A high-speed search engine pLink 2 with systematic evaluation for proteome-scale identification of cross-linked peptides. *Nat Commun* **10**, 3404, doi:10.1038/s41467-019-11337-z (2019).
61 Grimm, M., Zimniak, T., Kahraman, A. & Herzog, F. xVis: a web server for the schematic visualization and interpretation of crosslink-derived spatial restraints. *Nucleic Acids Res* **43**, W362-369, doi:10.1093/nar/gkv463 (2015).
62 Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J Mol Graph* **14**, 33-38, 27-38 (1996).
63 Schrödinger, L. The PyMOL Molecular Graphics System, Version 1.8. (2015).
64 Jensen, P. F., Rand, K. D. Hydrogen Exchange. In Hydrogen Exchange Mass Spectrometry of Proteins, D. D. Weis (Ed.). doi:10.1002/9781118703748.ch1. doi:10.1002/9781118703748.ch1 (2016).
65 Konermann, L., Pan, J. & Liu, Y. H. Hydrogen exchange mass spectrometry for studying protein structure and dynamics. *Chem Soc Rev* **40**, 1224-1234, doi:10.1039/c0cs00113a (2011).
66 Bai, Y., Milne, J. S., Mayne, L. & Englander, S. W. Primary structure effects on peptide group hydrogen exchange. *Proteins* **17**, 75-86, doi:10.1002/prot.340170110 (1993).
67 Wang, H., Rempel, D. L., Giblin, D., Frieden, C. & Gross, M. L. Peptide-Level Interactions between Proteins and Small-Molecule Drug Candidates by Two Hydrogen-Deuterium Exchange MS-Based Methods: The Example of Apolipoprotein E3. *Anal Chem* **89**, 10687-10695, doi:10.1021/acs.analchem.7b01121 (2017).
68 Hentze, N. & Mayer, M. P. Analyzing protein dynamics using hydrogen exchange mass spectrometry. *J Vis Exp*, doi:10.3791/50839 (2013).
69 Walters, B. T., Ricciuti, A., Mayne, L. & Englander, S. W. Minimizing back exchange in the hydrogen exchange-mass spectrometry experiment. *J Am Soc Mass Spectrom* **23**, 2132-2139, doi:10.1007/s13361-012-0476-x (2012).
70 Rey, M. *et al.* Nepenthesin from monkey cups for hydrogen/deuterium exchange mass spectrometry. *Mol Cell Proteomics* **12**, 464-472, doi:10.1074/mcp.M112.025221 (2013).

71 Nirudodhi, S. N., Sperry, J. B., Rouse, J. C. & Carroll, J. A. Application of Dual Protease Column for HDX-MS Analysis of Monoclonal Antibodies. *J Pharm Sci* **106**, 530-536, doi:10.1016/j.xphs.2016.10.023 (2017).
72 Chames, P., Van Regenmortel, M., Weiss, E. & Baty, D. Therapeutic antibodies: successes, limitations and hopes for the future. *Br J Pharmacol* **157**, 220-233, doi:10.1111/j.1476-5381.2009.00190.x (2009).
73 Irani, V. *et al.* Molecular properties of human IgG subclasses and their implications for designing therapeutic monoclonal antibodies against infectious diseases. *Mol Immunol* **67**, 171-182, doi:10.1016/j.molimm.2015.03.255 (2015).
74 Jakobovits, A., Amado, R. G., Yang, X., Roskos, L. & Schwab, G. From XenoMouse technology to panitumumab, the first fully human antibody product from transgenic mice. *Nat Biotechnol* **25**, 1134-1143, doi:10.1038/nbt1337 (2007).
75 Davies, A. M. & Sutton, B. J. Human IgG4: a structural perspective. *Immunol Rev* **268**, 139-159, doi:10.1111/imr.12349 (2015).
76 Murphy, K., Weaver, C., Mowat, A., Berg, L. & Chaplin, D. D. *Janeway's immunobiology*. 9th edition. edn, (Garland Science/Taylor & Francis Group, LLC, 2016).
77 Roux, K. H., Strelets, L. & Michaelsen, T. E. Flexibility of human IgG subclasses. *J Immunol* **159**, 3372-3382 (1997).
78 Schroeder, H. W., Jr. & Cavacini, L. Structure and function of immunoglobulins. *J Allergy Clin Immunol* **125**, S41-52, doi:10.1016/j.jaci.2009.09.046 (2010).
79 Krishnamurthy, A. & Jimeno, A. Bispecific antibodies for cancer therapy: A review. *Pharmacol Ther* **185**, 122-134, doi:10.1016/j.pharmthera.2017.12.002 (2018).
80 Trivedi, A. *et al.* Clinical Pharmacology and Translational Aspects of Bispecific Antibodies. *Clin Transl Sci* **10**, 147-162, doi:10.1111/cts.12459 (2017).
81 Diamantis, N. & Banerji, U. Antibody-drug conjugates--an emerging class of cancer treatment. *Br J Cancer* **114**, 362-367, doi:10.1038/bjc.2015.435 (2016).
82 Liu-Shin, L., Fung, A., Malhotra, A. & Ratnaswamy, G. Influence of disulfide bond isoforms on drug conjugation sites in cysteine-linked IgG2 antibody-drug conjugates. *MAbs* **10**, 583-595, doi:10.1080/19420862.2018.1440165 (2018).
83 Perez, H. L. *et al.* Antibody-drug conjugates: current status and future directions. *Drug Discov Today* **19**, 869-881, doi:10.1016/j.drudis.2013.11.004 (2014).
84 Kaplon, H. & Reichert, J. M. Antibodies to watch in 2018. *MAbs* **10**, 183-203, doi:10.1080/19420862.2018.1415671 (2018).
85 Michaelsen, T. E., Frangione, B. & Franklin, E. C. Primary structure of the "hinge" region of human IgG3. Probable quadruplication of a 15-amino acid residue basic unit. *J Biol Chem* **252**, 883-889 (1977).
86 Vidarsson, G., Dekkers, G. & Rispens, T. IgG subclasses and allotypes: from structure to effector functions. *Front Immunol* **5**, 520, doi:10.3389/fimmu.2014.00520 (2014).
87 Brekke, O. H., Michaelsen, T. E. & Sandlie, I. The structural requirements for complement activation by IgG: does it hinge on the hinge? *Immunol Today* **16**, 85-90, doi:10.1016/0167-5699(95)80094-8 (1995).
88 Rayner, L. E. *et al.* The Fab conformations in the solution structure of human immunoglobulin G4 (IgG4) restrict access to its Fc region: implications for functional activity. *J Biol Chem* **289**, 20740-20756, doi:10.1074/jbc.M114.572404 (2014).
89 Ahdash, Z., Pyle, E. & Politis, A. Hybrid Mass Spectrometry: Towards Characterization of Protein Conformational States. *Trends Biochem Sci* **41**, 650-653, doi:10.1016/j.tibs.2016.04.008 (2016).
90 Goth, M. & Pagel, K. Ion mobility-mass spectrometry as a tool to investigate protein-ligand interactions. *Analytical and bioanalytical chemistry* **409**, 4305-4310, doi:10.1007/s00216-017-0384-9 (2017).

91 Hernandez, H. & Robinson, C. V. Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nat Protoc* **2**, 715-726, doi:10.1038/nprot.2007.73 (2007).
92 Konijnenberg, A., Butterer, A. & Sobott, F. Native ion mobility-mass spectrometry and related methods in structural biology. *Biochim Biophys Acta* **1834**, 1239-1256, doi:10.1016/j.bbapap.2012.11.013 (2013).
93 Scarff, C. A., Thalassinos, K., Hilton, G. R. & Scrivens, J. H. Travelling wave ion mobility mass spectrometry studies of protein structure: biological significance and comparison with X-ray crystallography and nuclear magnetic resonance spectroscopy measurements. *Rapid Commun Mass Spectrom* **22**, 3297-3304, doi:10.1002/rcm.3737 (2008).
94 Sharon, M. & Robinson, C. V. The role of mass spectrometry in structure elucidation of dynamic protein complexes. *Annu Rev Biochem* **76**, 167-193, doi:10.1146/annurev.biochem.76.061005.090816 (2007).
95 Watanabe, Y. *et al.* Signature of Antibody Domain Exchange by Native Mass Spectrometry and Collision-Induced Unfolding. *Anal Chem*, doi:10.1021/acs.analchem.8b00573 (2018).
96 Eschweiler, J. D., Farrugia, M. A., Dixit, S. M., Hausinger, R. P. & Ruotolo, B. T. A Structural Model of the Urease Activation Complex Derived from Ion Mobility-Mass Spectrometry and Integrative Modeling. *Structure* **26**, 599-606 e593, doi:10.1016/j.str.2018.03.001 (2018).
97 Hall, Z., Politis, A. & Robinson, C. V. Structural modeling of heteromeric protein complexes from disassembly pathways and ion mobility-mass spectrometry. *Structure* **20**, 1596-1609, doi:10.1016/j.str.2012.07.001 (2012).
98 Kebarle, P. & Verkerk, U. H. Electrospray: from ions in solution to ions in the gas phase, what we know now. *Mass Spectrom Rev* **28**, 898-917, doi:10.1002/mas.20247 (2009).
99 Wilm, M. S. & Mann, M. Electrospray and Taylor-Cone theory, Dole's beam of macromolecules at last? *International Journal of Mass Spectrometry and Ion Processes* **136**, 167-180, doi:https://doi.org/10.1016/0168-1176(94)04024-9 (1994).
100 Porrini, M. *et al.* Compaction of Duplex Nucleic Acids upon Native Electrospray Mass Spectrometry. *ACS Cent Sci* **3**, 454-461, doi:10.1021/acscentsci.7b00084 (2017).
101 Hogan, C. J., Jr., Ruotolo, B. T., Robinson, C. V. & Fernandez de la Mora, J. Tandem differential mobility analysis-mass spectrometry reveals partial gas-phase collapse of the GroEL complex. *J Phys Chem B* **115**, 3614-3621, doi:10.1021/jp109172k (2011).
102 Konermann, L. Molecular Dynamics Simulations on Gas-Phase Proteins with Mobile Protons: Inclusion of All-Atom Charge Solvation. *J Phys Chem B* **121**, 8102-8112, doi:10.1021/acs.jpcb.7b05703 (2017).
103 Bush, M. F. *et al.* Collision cross sections of proteins and their complexes: a calibration framework and database for gas-phase structural biology. *Anal Chem* **82**, 9557-9565, doi:10.1021/ac1022953 (2010).
104 Fiser, A., Do, R. K. & Sali, A. Modeling of loops in protein structures. *Protein Sci* **9**, 1753-1773, doi:10.1110/ps.9.9.1753 (2000).
105 Shen, M. Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci* **15**, 2507-2524, doi:10.1110/ps.062416606 (2006).
106 Berendsen, H. J. C., Vanderspoel, D. & Vandrunen, R. Gromacs - a Message-Passing Parallel Molecular-Dynamics Implementation. *Comput Phys Commun* **91**, 43-56, doi:Doi 10.1016/0010-4655(95)00042-E (1995).
107 MacKerell, A. D. *et al.* All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B* **102**, 3586-3616, doi:10.1021/jp973084f (1998).

108 Russel, D. *et al.* Putting the pieces together: integrative modeling platform software for structure determination of macromolecular assemblies. *PLoS Biol* **10**, e1001244, doi:10.1371/journal.pbio.1001244 (2012).
109 Tan, K. P., Nguyen, T. B., Patel, S., Varadarajan, R. & Madhusudhan, M. S. Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. *Nucleic Acids Res* **41**, W314-321, doi:10.1093/nar/gkt503 (2013).
110 Young, M. N. & Bleiholder, C. Molecular Structures and Momentum Transfer Cross Sections: The Influence of the Analyte Charge Distribution. *J Am Soc Mass Spectrom* **28**, 619-627, doi:10.1007/s13361-017-1605-3 (2017).
111 Bleiholder, C., Contreras, S. & Bowers, M. T. A novel projection approximation algorithm for the fast and accurate computation of molecular collision cross sections (IV). Application to polypeptides. *Int J Mass Spectrom* **354**, 275-280, doi:10.1016/j.ijms.2013.06.011 (2013).
112 Harris, L. J., Skaletsky, E. & McPherson, A. Crystallographic structure of an intact IgG1 monoclonal antibody. *J Mol Biol* **275**, 861-872, doi:10.1006/jmbi.1997.1508 (1998).
113 Saphire, E. O. *et al.* Crystal structure of a neutralizing human IGG against HIV-1: a template for vaccine design. *Science* **293**, 1155-1159, doi:10.1126/science.1061692 (2001).
114 Harris, L. J., Larson, S. B., Hasel, K. W. & McPherson, A. Refined structure of an intact IgG2a monoclonal antibody. *Biochemistry* **36**, 1581-1597, doi:10.1021/bi962514+ (1997).
115 Scapin, G. *et al.* Structure of full-length human anti-PD1 therapeutic IgG4 antibody pembrolizumab. *Nat Struct Mol Biol* **22**, 953-958, doi:10.1038/nsmb.3129 (2015).
116 Gregory, L. *et al.* The solution conformations of the subclasses of human IgG deduced from sedimentation and small angle X-ray scattering studies. *Mol Immunol* **24**, 821-829 (1987).
117 Phillips, M. L., Tao, M. H., Morrison, S. L. & Schumaker, V. N. Human/mouse chimeric monoclonal antibodies with human IgG1, IgG2, IgG3 and IgG4 constant domains: electron microscopic and hydrodynamic characterization. *Mol Immunol* **31**, 1201-1210 (1994).
118 Carrasco, B. *et al.* Crystallohydrodynamics for solving the hydration problem for multi-domain proteins: open physiological conformations for human IgG. *Biophys Chem* **93**, 181-196 (2001).
119 Rayner, L. E. *et al.* The solution structures of two human IgG1 antibodies show conformational stability and accommodate their C1q and FcgammaR ligands. *J Biol Chem* **290**, 8420-8438, doi:10.1074/jbc.M114.631002 (2015).
120 Tian, X. *et al.* In-depth analysis of subclass-specific conformational preferences of IgG antibodies. *IUCrJ* **2**, 9-18, doi:10.1107/S205225251402209X (2015).
121 Bruhns, P. *et al.* Specificity and affinity of human Fcgamma receptors and their polymorphic variants for human IgG subclasses. *Blood* **113**, 3716-3725, doi:10.1182/blood-2008-09-179754 (2009).
122 Pascal, B. D. *et al.* HDX workbench: software for the analysis of H/D exchange MS data. *J Am Soc Mass Spectrom* **23**, 1512-1521, doi:10.1007/s13361-012-0419-6 (2012).
123 Lou, X. *et al.* Deuteration distribution estimation with improved sequence coverage for HX/MS experiments. *Bioinformatics* **26**, 1535-1541, doi:10.1093/bioinformatics/btq165 (2010).
124 Palmblad, M., Buijs, J. & Hakansson, P. Automatic analysis of hydrogen/deuterium exchange mass spectra of peptides and proteins using calculations of isotopic distributions. *J Am Soc Mass Spectrom* **12**, 1153-1162, doi:10.1016/S1044-0305(01)00301-4 (2001).

125 Slysz, G. W. *et al.* Hydra: software for tailored processing of H/D exchange data from MS or tandem MS analyses. *BMC Bioinformatics* **10**, 162, doi:10.1186/1471-2105-10-162 (2009).
126 Woods, V. L., Jr. & Hamuro, Y. High resolution, high-throughput amide deuterium exchange-mass spectrometry (DXMS) determination of protein binding site structure and dynamics: utility in pharmaceutical design. *J Cell Biochem Suppl* **Suppl 37**, 89-98 (2001).
127 Weis, D. D., Engen, J. R. & Kass, I. J. Semi-automated data processing of hydrogen exchange mass spectra using HX-Express. *J Am Soc Mass Spectrom* **17**, 1700-1703, doi:10.1016/j.jasms.2006.07.025 (2006).
128 Pascal, B. D. *et al.* The Deuterator: software for the determination of backbone amide deuterium levels from H/D exchange MS data. *BMC Bioinformatics* **8**, 156, doi:10.1186/1471-2105-8-156 (2007).
129 Nikamanon, P., Pun, E., Chou, W., Koter, M. D. & Gershon, P. D. "TOF2H": a precision toolbox for rapid, high density/high coverage hydrogen-deuterium exchange mass spectrometry via an LC-MALDI approach, covering the data pipeline from spectral acquisition to HDX rate analysis. *BMC Bioinformatics* **9**, 387, doi:10.1186/1471-2105-9-387 (2008).
130 Pascal, B. D., Chalmers, M. J., Busby, S. A. & Griffin, P. R. HD desktop: an integrated platform for the analysis and visualization of H/D exchange data. *J Am Soc Mass Spectrom* **20**, 601-610, doi:10.1016/j.jasms.2008.11.019 (2009).
131 Kavan, D. & Man, P. MSTools-Web based application for visualization and presentation of HXMS data. *Int J Mass Spectrom* **302**, 53-58, doi:10.1016/j.ijms.2010.07.030 (2011).
132 Liu, S. *et al.* HDX-analyzer: a novel package for statistical analysis of protein structure dynamics. *BMC Bioinformatics* **12 Suppl 1**, S43, doi:10.1186/1471-2105-12-S1-S43 (2011).
133 Kan, Z. Y., Mayne, L., Chetty, P. S. & Englander, S. W. ExMS: data analysis for HX-MS experiments. *J Am Soc Mass Spectrom* **22**, 1906-1915, doi:10.1007/s13361-011-0236-3 (2011).
134 Miller, D. E., Prasannan, C. B., Villar, M. T., Fenton, A. W. & Artigues, A. HDXFinder: automated analysis and data reporting of deuterium/hydrogen exchange mass spectrometry. *J Am Soc Mass Spectrom* **23**, 425-429, doi:10.1007/s13361-011-0234-5 (2012).
135 Rey, M. *et al.* Mass spec studio for integrative structural biology. *Structure* **22**, 1538-1548, doi:10.1016/j.str.2014.08.013 (2014).
136 Lindner, R. *et al.* Hexicon 2: automated processing of hydrogen-deuterium exchange mass spectrometry data with improved deuteration distribution estimation. *J Am Soc Mass Spectrom* **25**, 1018-1028, doi:10.1007/s13361-014-0850-y (2014).
137 Hourdel, V. *et al.* MEMHDX: an interactive tool to expedite the statistical validation and visualization of large HDX-MS datasets. *Bioinformatics* **32**, 3413-3419, doi:10.1093/bioinformatics/btw420 (2016).
138 Lau, A. M. C., Ahdash, Z., Martens, C. & Politis, A. Deuteros: software for rapid analysis and visualization of data from differential hydrogen deuterium exchange-mass spectrometry. *Bioinformatics*, doi:10.1093/bioinformatics/btz022 (2019).
139 Bouyssie, D. *et al.* HDX-Viewer: interactive 3D visualization of hydrogen-deuterium exchange data. *Bioinformatics*, doi:10.1093/bioinformatics/btz550 (2019).
140 Houde, D., Berkowitz, S. A. & Engen, J. R. The utility of hydrogen/deuterium exchange mass spectrometry in biopharmaceutical comparability studies. *J Pharm Sci* **100**, 2071-2086, doi:10.1002/jps.22432 (2011).
141 Martens, C. *et al.* Direct protein-lipid interactions shape the conformational landscape of secondary transporters. *Nature Communications* **9**, 4151, doi:10.1038/s41467-018-06704-1 (2018).

142 Reading, E. *et al.* Interrogating membrane protein conformational dynamics within native lipid compositions. *Angew Chem Int Ed Engl*, doi:10.1002/anie.201709657 (2017).
143 Killoran, R. C., Sowole, M. A., Halim, M. A., Konermann, L. & Choy, W. Y. Conformational characterization of the intrinsically disordered protein Chibby: Interplay between structural elements in target recognition. *Protein Sci* **25**, 1420-1429, doi:10.1002/pro.2936 (2016).
144 Suchanova, B. & Tuma, R. Folding and assembly of large macromolecular complexes monitored by hydrogen-deuterium exchange and mass spectrometry. *Microb Cell Fact* **7**, 12, doi:10.1186/1475-2859-7-12 (2008).
145 Xiao, K. *et al.* Revealing the architecture of protein complexes by an orthogonal approach combining HDXMS, CXMS, and disulfide trapping. *Nat Protoc* **13**, 1403-1428, doi:10.1038/nprot.2018.037 (2018).
146 Masson, G. R. *et al.* Recommendations for performing, interpreting and reporting hydrogen deuterium exchange mass spectrometry (HDX-MS) experiments. *Nat Methods* **16**, 595-602, doi:10.1038/s41592-019-0459-y (2019).
147 Rice, T. K., Schork, N. J. & Rao, D. C. in *Genetic Dissection of Complex Traits Advances in Genetics* 293-308 (2008).
148 Majumdar, R. *et al.* Effects of salts from the Hofmeister series on the conformational stability, aggregation propensity, and local flexibility of an IgG1 monoclonal antibody. *Biochemistry* **52**, 3376-3389, doi:10.1021/bi400232p (2013).
149 Pan, L. Y., Salas-Solano, O. & Valliere-Douglass, J. F. Conformation and dynamics of interchain cysteine-linked antibody-drug conjugates as revealed by hydrogen/deuterium exchange mass spectrometry. *Anal Chem* **86**, 2657-2664, doi:10.1021/ac404003q (2014).
150 Morgan, C. R. *et al.* Conformational transitions in the membrane scaffold protein of phospholipid bilayer nanodiscs. *Mol Cell Proteomics* **10**, M111 010876, doi:10.1074/mcp.M111.010876 (2011).
151 Ahdash, Z. *et al.* HDX-MS reveals nucleotide-dependent, anti-correlated opening and closure of SecA and SecY channels of the bacterial translocon. *Elife* **8**, doi:10.7554/eLife.47402 (2019).
152 Habibi, Y., Uggowitzer, K. A., Issak, H. & Thibodeaux, C. J. Insights into the Dynamic Structural Properties of a Lanthipeptide Synthetase using Hydrogen-Deuterium Exchange Mass Spectrometry. *J Am Chem Soc* **141**, 14661-14672, doi:10.1021/jacs.9b06020 (2019).
153 Terral, G. *et al.* Epitope characterization of anti-JAM-A antibodies using orthogonal mass spectrometry and surface plasmon resonance approaches. *MAbs* **9**, 1317-1326, doi:10.1080/19420862.2017.1380762 (2017).
154 van Erp, P. B. G. *et al.* Conformational Dynamics of DNA Binding and Cas3 Recruitment by the CRISPR RNA-Guided Cascade Complex. *ACS Chem Biol* **13**, 481-490, doi:10.1021/acschembio.7b00649 (2018).
155 Rochel, N. *et al.* Recurrent activating mutations of PPARgamma associated with luminal bladder tumors. *Nat Commun* **10**, 253, doi:10.1038/s41467-018-08157-y (2019).
156 Guilvout, I. *et al.* Prepore Stability Controls Productive Folding of the BAM-independent Multimeric Outer Membrane Secretin PulD. *J Biol Chem* **292**, 328-338, doi:10.1074/jbc.M116.759498 (2017).
157 Faull, S. V. *et al.* Structural basis of Cullin 2 RING E3 ligase regulation by the COP9 signalosome. *Nat Commun* **10**, 3814, doi:10.1038/s41467-019-11772-y (2019).
158 Ciechanover, A. The ubiquitin-proteasome proteolytic pathway. *Cell* **79**, 13-21, doi:10.1016/0092-8674(94)90396-4 (1994).
159 Rape, M. Ubiquitylation at the crossroads of development and disease. *Nat Rev Mol Cell Biol* **19**, 59-70, doi:10.1038/nrm.2017.83 (2018).

160 Asaoka, T. & Ikeda, F. New Insights into the Role of Ubiquitin Networks in the Regulation of Antiapoptosis Pathways. *Int Rev Cell Mol Biol* **318**, 121-158, doi:10.1016/bs.ircmb.2015.05.003 (2015).
161 Platta, H. W., Thoms, S., Kunau, W. H. & Erdmann, R. in *Molecular Machines Involved in Protein Transport across Cellular Membranes The Enzymes* 541-572 (2007).
162 Stewart, M. D., Ritterhoff, T., Klevit, R. E. & Brzovic, P. S. E2 enzymes: more than just middle men. *Cell Res* **26**, 423-440, doi:10.1038/cr.2016.35 (2016).
163 Guerra, D. D. & Callis, J. Ubiquitin on the move: the ubiquitin modification system plays diverse roles in the regulation of endoplasmic reticulum- and plasma membrane-localized proteins. *Plant Physiol* **160**, 56-64, doi:10.1104/pp.112.199869 (2012).
164 Siepmann, T. J., Bohnsack, R. N., Tokgoz, Z., Baboshina, O. V. & Haas, A. L. Protein interactions within the N-end rule ubiquitin ligation pathway. *J Biol Chem* **278**, 9448-9457, doi:10.1074/jbc.M211240200 (2003).
165 Komander, D. & Rape, M. The ubiquitin code. *Annu Rev Biochem* **81**, 203-229, doi:10.1146/annurev-biochem-060310-170328 (2012).
166 Tanaka, K. The proteasome: overview of structure and functions. *Proc Jpn Acad Ser B Phys Biol Sci* **85**, 12-36, doi:10.2183/pjab.85.12 (2009).
167 Bhat, K. P. & Greer, S. F. Proteolytic and non-proteolytic roles of ubiquitin and the ubiquitin proteasome system in transcriptional regulation. *Biochim Biophys Acta* **1809**, 150-155, doi:10.1016/j.bbagrm.2010.11.006 (2011).
168 Yoshimura, T. *et al.* Molecular characterization of the "26S" proteasome complex from rat liver. *J Struct Biol* **111**, 200-211 (1993).
169 Sarikas, A., Hartmann, T. & Pan, Z. Q. The cullin protein family. *Genome Biol* **12**, 220, doi:10.1186/gb-2011-12-4-220 (2011).
170 Petroski, M. D. & Deshaies, R. J. Function and regulation of cullin-RING ubiquitin ligases. *Nat Rev Mol Cell Biol* **6**, 9-20, doi:10.1038/nrm1547 (2005).
171 Soucy, T. A. *et al.* An inhibitor of NEDD8-activating enzyme as a new approach to treat cancer. *Nature* **458**, 732-736, doi:10.1038/nature07884 (2009).
172 Yongchao, Z. & Yi, S. Cullin-RING Ligases as Attractive Anti-cancer Targets. *Current Pharmaceutical Design* **19**, 3215-3225, doi:http://dx.doi.org/10.2174/1381612811319999030 (2013).
173 Deshaies, R. J. & Joazeiro, C. A. RING domain E3 ubiquitin ligases. *Annu Rev Biochem* **78**, 399-434, doi:10.1146/annurev.biochem.78.101807.093809 (2009).
174 Zheng, N. *et al.* Structure of the Cul1–Rbx1–Skp1–F boxSkp2 SCF ubiquitin ligase complex. *Nature* **416**, 703-709, doi:10.1038/416703a (2002).
175 Goldenberg, S. J. *et al.* Structure of the Cand1-Cul1-Roc1 complex reveals regulatory mechanisms for the assembly of the multisubunit cullin-dependent ubiquitin ligases. *Cell* **119**, 517-528, doi:10.1016/j.cell.2004.10.019 (2004).
176 Calabrese, M. F. *et al.* A RING E3–substrate complex poised for ubiquitin-like protein transfer: structural insights into cullin-RING ligases. *Nature Structural & Molecular Biology* **18**, 947-949, doi:10.1038/nsmb.2086 (2011).
177 Scott, D. C., Monda, J. K., Bennett, E. J., Harper, J. W. & Schulman, B. A. N-Terminal Acetylation Acts as an Avidity Enhancer Within an Interconnected Multiprotein Complex. *Science* **334**, 674, doi:10.1126/science.1209307 (2011).
178 Duda, David M. *et al.* Structure of a Glomulin-RBX1-CUL1 Complex: Inhibition of a RING E3 Ligase through Masking of Its E2-Binding Surface. *Molecular Cell* **47**, 371-382, doi:https://doi.org/10.1016/j.molcel.2012.05.044 (2012).
179 Scott, D. C. *et al.* Structure of a RING E3 trapped in action reveals ligation mechanism for the ubiquitin-like protein NEDD8. *Cell* **157**, 1671-1684, doi:10.1016/j.cell.2014.04.037 (2014).

180 Scott, D. C. *et al.* Blocking an N-terminal acetylation–dependent protein interaction inhibits an E3 ligase. *Nature Chemical Biology* **13**, 850, doi:10.1038/nchembio.2386

https://www.nature.com/articles/nchembio.2386#supplementary-information (2017).

181 Nguyen, Henry C., Yang, H., Fribourgh, Jennifer L., Wolfe, Leslie S. & Xiong, Y. Insights into Cullin-RING E3 Ubiquitin Ligase Recruitment: Structure of the VHL-EloBC-Cul2 Complex. *Structure* **23**, 441-449, doi:https://doi.org/10.1016/j.str.2014.12.014 (2015).

182 Cardote, T. A. F., Gadd, M. S. & Ciulli, A. Crystal Structure of the Cul2-Rbx1-EloBC-VHL Ubiquitin Ligase Complex. *Structure* **25**, 901-911 e903, doi:10.1016/j.str.2017.04.009 (2017).

183 Errington, Wesley J. *et al.* Adaptor Protein Self-Assembly Drives the Control of a Cullin-RING Ubiquitin Ligase. *Structure* **20**, 1141-1153, doi:https://doi.org/10.1016/j.str.2012.04.009 (2012).

184 Canning, P. *et al.* Structural Basis for Cul3 Protein Assembly with the BTB-Kelch Family of E3 Ubiquitin Ligases. *Journal of Biological Chemistry* **288**, 7803-7814 (2013).

185 Gao, C., Pallett, M. A., Croll, T. I., Smith, G. L. & Graham, S. C. Molecular basis of cullin-3 (Cul3) ubiquitin ligase subversion by vaccinia virus protein A55. *Journal of Biological Chemistry* **294**, 6416-6429, doi:10.1074/jbc.RA118.006561 (2019).

186 Angers, S. *et al.* Molecular architecture and assembly of the DDB1–CUL4A ubiquitin ligase machinery. *Nature* **443**, 590-593, doi:10.1038/nature05175 (2006).

187 Fischer, Eric S. *et al.* The Molecular Basis of CRL4DDB2/CSA Ubiquitin Ligase Architecture, Targeting, and Activation. *Cell* **147**, 1024-1039, doi:https://doi.org/10.1016/j.cell.2011.10.035 (2011).

188 Duda, D. M. *et al.* Structural insights into NEDD8 activation of cullin-RING ligases: conformational control of conjugation. *Cell* **134**, 995-1006, doi:10.1016/j.cell.2008.07.022 (2008).

189 Kim, Y. K. *et al.* Structural basis of intersubunit recognition in elongin BC-cullin 5-SOCS box ubiquitin-protein ligase complexes. *Acta Crystallographica Section D* **69**, 1587-1597, doi:doi:10.1107/S0907444913011220 (2013).

190 Guo, Y. *et al.* Structural basis for hijacking CBF-β and CUL5 E3 ligase complex by HIV-1 Vif. *Nature* **505**, 229, doi:10.1038/nature12884 (2014).

191 Feldman, R. M. R., Correll, C. C., Kaplan, K. B. & Deshaies, R. J. A Complex of Cdc4p, Skp1p, and Cdc53p/Cullin Catalyzes Ubiquitination of the Phosphorylated CDK Inhibitor Sic1p. *Cell* **91**, 221-230, doi:https://doi.org/10.1016/S0092-8674(00)80404-3 (1997).

192 Skowyra, D., Craig, K. L., Tyers, M., Elledge, S. J. & Harper, J. W. F-Box Proteins Are Receptors that Recruit Phosphorylated Substrates to the SCF Ubiquitin-Ligase Complex. *Cell* **91**, 209-219, doi:https://doi.org/10.1016/S0092-8674(00)80403-1 (1997).

193 Bulatov, E. & Ciulli, A. Targeting Cullin-RING E3 ubiquitin ligases for drug discovery: structure, assembly and small-molecule modulation. *Biochem J* **467**, 365-386, doi:10.1042/BJ20141450 (2015).

194 Wang, G., Chan, C. H., Gao, Y. & Lin, H. K. Novel roles of Skp2 E3 ligase in cellular senescence, cancer progression, and metastasis. *Chin J Cancer* **31**, 169-177, doi:10.5732/cjc.011.10319 (2012).

195 Davis, R. J., Welcker, M. & Clurman, B. E. Tumor suppression by the Fbw7 ubiquitin ligase: mechanisms and opportunities. *Cancer Cell* **26**, 455-464, doi:10.1016/j.ccell.2014.09.013 (2014).

196 Cai, W. & Yang, H. The structure and regulation of Cullin 2 based E3 ubiquitin ligases and their biological functions. *Cell Div* **11**, 7, doi:10.1186/s13008-016-0020-7 (2016).

197 Jackson, S. & Xiong, Y. CRL4s: the CUL4-RING E3 ubiquitin ligases. *Trends Biochem Sci* **34**, 562-570, doi:10.1016/j.tibs.2009.07.002 (2009).

198 Wang, S. *et al.* Atlas on substrate recognition subunits of CRL2 E3 ligases. *Oncotarget* **7**, 46707-46716, doi:10.18632/oncotarget.8732 (2016).
199 Otrock, Z. K., Mahfouz, R. A. R., Makarem, J. A. & Shamseddine, A. I. Understanding the biology of angiogenesis: Review of the most important molecular mechanisms. *Blood Cells, Molecules, and Diseases* **39**, 212-220, doi:https://doi.org/10.1016/j.bcmd.2007.04.001 (2007).
200 Sakamoto, K. M. *et al.* Protacs: Chimeric molecules that target proteins to the Skp1–Cullin–F box complex for ubiquitination and degradation. *Proceedings of the National Academy of Sciences* **98**, 8554, doi:10.1073/pnas.141230798 (2001).
201 Zou, Y., Ma, D. & Wang, Y. The PROTAC technology in drug development. *Cell Biochem Funct* **37**, 21-30, doi:10.1002/cbf.3369 (2019).
202 Zhang, D. *et al.* Targeted degradation of proteins by small molecules: a novel tool for functional proteomics. *Comb Chem High Throughput Screen* **7**, 689-697, doi:10.2174/1386207043328364 (2004).
203 Bosu, D. R. & Kipreos, E. T. Cullin-RING ubiquitin ligases: global regulation and activation cycles. *Cell Div* **3**, 7, doi:10.1186/1747-1028-3-7 (2008).
204 Merlet, J., Burger, J., Gomes, J. E. & Pintard, L. Regulation of cullin-RING E3 ubiquitin-ligases by neddylation and dimerization. *Cell Mol Life Sci* **66**, 1924-1938, doi:10.1007/s00018-009-8712-7 (2009).
205 Yamoah, K. *et al.* Autoinhibitory regulation of SCF-mediated ubiquitination by human cullin 1's C-terminal tail. *Proceedings of the National Academy of Sciences* **105**, 12230, doi:10.1073/pnas.0806155105 (2008).
206 Chua, Y. S., Boh, B. K., Ponyeam, W. & Hagen, T. Regulation of cullin RING E3 ubiquitin ligases by CAND1 in vivo. *PLoS One* **6**, e16071, doi:10.1371/journal.pone.0016071 (2011).
207 Soucy, T. A., Dick, L. R., Smith, P. G., Milhollen, M. A. & Brownell, J. E. The NEDD8 Conjugation Pathway and Its Relevance in Cancer Biology and Therapy. *Genes Cancer* **1**, 708-716, doi:10.1177/1947601910382898 (2010).
208 Kurz, T. *et al.* The conserved protein DCN-1/Dcn1p is required for cullin neddylation in C. elegans and S. cerevisiae. *Nature* **435**, 1257-1261, doi:10.1038/nature03662 (2005).
209 Duda, D. M. *et al.* Structural regulation of cullin-RING ubiquitin ligase complexes. *Curr Opin Struct Biol* **21**, 257-264, doi:10.1016/j.sbi.2011.01.003 (2011).
210 Maeda, Y. *et al.* CUL2 is required for the activity of hypoxia-inducible factor and vasculogenesis. *J Biol Chem* **283**, 16084-16092, doi:10.1074/jbc.M710223200 (2008).
211 Nguyen, H. C., Yang, H., Fribourgh, J. L., Wolfe, L. S. & Xiong, Y. Insights into Cullin-RING E3 ubiquitin ligase recruitment: structure of the VHL-EloBC-Cul2 complex. *Structure* **23**, 441-449, doi:10.1016/j.str.2014.12.014 (2015).
212 Pause, A. *et al.* The von Hippel-Lindau tumor-suppressor gene product forms a stable complex with human CUL-2, a member of the Cdc53 family of proteins. *Proc Natl Acad Sci U S A* **94**, 2156-2161 (1997).
213 Deshaies, R. J. Protein degradation: Prime time for PROTACs. *Nat Chem Biol* **11**, 634-635, doi:10.1038/nchembio.1887 (2015).
214 Soares, P. *et al.* Group-based optimization of potent and cell-active inhibitors of the von Hippel-Lindau (VHL) E3 ubiquitin ligase: structure-activity relationships leading to the chemical probe (2S,4R)-1-((S)-2-(1-cyanocyclopropanecarboxamido)-3,3-dimethylbutanoyl)-4-hydroxy -N-(4-(4-methylthiazol-5-yl)benzyl)pyrrolidine-2-carboxamide (VH298). *J Med Chem*, doi:10.1021/acs.jmedchem.7b00675 (2017).
215 Wada, H., Yeh, E. T. & Kamitani, T. Identification of NEDD8-conjugation site in human cullin-2. *Biochem Biophys Res Commun* **257**, 100-105, doi:10.1006/bbrc.1999.0339 (1999).
216 Enchev, R. I., Schreiber, A., Beuron, F. & Morris, E. P. Structural insights into the COP9 signalosome and its common architecture with the 26S proteasome lid and eIF3. *Structure* **18**, 518-527, doi:10.1016/j.str.2010.02.008 (2010).

217 Lingaraju, G. M. *et al.* Crystal structure of the human COP9 signalosome. *Nature* **512**, 161-165, doi:10.1038/nature13566 (2014).
218 Sharon, M. *et al.* Symmetrical modularity of the COP9 signalosome complex suggests its multifunctionality. *Structure* **17**, 31-40, doi:10.1016/j.str.2008.10.012 (2009).
219 Enchev, R. I. *et al.* Structural basis for a reciprocal regulation between SCF and CSN. *Cell Rep* **2**, 616-627, doi:10.1016/j.celrep.2012.08.019 (2012).
220 Pick, E., Hofmann, K. & Glickman, M. H. PCI complexes: Beyond the proteasome, CSN, and eIF3 Troika. *Mol Cell* **35**, 260-264, doi:10.1016/j.molcel.2009.07.009 (2009).
221 Rozen, S. *et al.* CSNAP Is a Stoichiometric Subunit of the COP9 Signalosome. *Cell Rep* **13**, 585-598, doi:10.1016/j.celrep.2015.09.021 (2015).
222 Mosadeghi, R. *et al.* Structural and kinetic analysis of the COP9-Signalosome activation and the cullin-RING ubiquitin ligase deneddylation cycle. *Elife* **5**, doi:10.7554/eLife.12102 (2016).
223 Cavadini, S. *et al.* Cullin-RING ubiquitin E3 ligase regulation by the COP9 signalosome. *Nature* **531**, 598-603, doi:10.1038/nature17416 (2016).
224 Cope, G. A. *et al.* Role of predicted metalloprotease motif of Jab1/Csn5 in cleavage of Nedd8 from Cul1. *Science* **298**, 608-611, doi:10.1126/science.1075901 (2002).
225 Echalier, A. *et al.* Insights into the regulation of the human COP9 signalosome catalytic subunit, CSN5/Jab1. *Proc Natl Acad Sci U S A* **110**, 1273-1278, doi:10.1073/pnas.1209345110 (2013).
226 Bennett, E. J., Rush, J., Gygi, S. P. & Harper, J. W. Dynamics of Cullin-RING Ubiquitin Ligase Network Revealed by Systematic Quantitative Proteomics. *Cell* **143**, 951-965, doi:10.1016/j.cell.2010.11.017 (2010).
227 Emberley, E. D., Mosadeghi, R. & Deshaies, R. J. Deconjugation of Nedd8 from Cul1 is directly regulated by Skp1-F-box and substrate, and the COP9 signalosome inhibits deneddylated SCF by a noncatalytic mechanism. *J Biol Chem* **287**, 29679-29689, doi:10.1074/jbc.M112.352484 (2012).
228 Ohta, T., Michel, J. J., Schottelius, A. J. & Xiong, Y. ROC1, a homolog of APC11, represents a family of cullin partners with an associated ubiquitin ligase activity. *Mol Cell* **3**, 535-541 (1999).
229 Li, L. *et al.* Hypoxia-inducible factor linked to differential kidney cancer risk seen with type 2A and type 2B VHL mutations. *Mol Cell Biol* **27**, 5381-5392, doi:10.1128/MCB.00282-07 (2007).
230 Sari, D. *et al.* The MultiBac Baculovirus/Insect Cell Expression Vector System for Producing Complex Protein Biologics. *Adv Exp Med Biol* **896**, 199-215, doi:10.1007/978-3-319-27216-0_13 (2016).
231 Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* **180**, 519-530, doi:10.1016/j.jsb.2012.09.006 (2012).
232 Zheng, S. Q. *et al.* MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat Methods* **14**, 331-332, doi:10.1038/nmeth.4193 (2017).
233 Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J Struct Biol* **192**, 216-221, doi:10.1016/j.jsb.2015.08.008 (2015).
234 Stark, H. & Chari, A. Sample preparation of biological macromolecular assemblies for the determination of high-resolution structures by cryo-electron microscopy. *Microscopy (Oxf)* **65**, 23-34, doi:10.1093/jmicro/dfv367 (2016).
235 Zivanov, J. *et al.* New tools for automated high-resolution cryo-EM structure determination in RELION-3. *Elife* **7**, doi:10.7554/eLife.42166 (2018).
236 Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat Methods* **11**, 63-65, doi:10.1038/nmeth.2727 (2014).

237 Webb, B. & Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Protein Sci* **86**, 2 9 1-2 9 37, doi:10.1002/cpps.20 (2016).
238 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612, doi:10.1002/jcc.20084 (2004).
239 Trabuco, L. G., Villa, E., Schreiner, E., Harrison, C. B. & Schulten, K. Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. *Methods* **49**, 174-180, doi:10.1016/j.ymeth.2009.04.005 (2009).
240 Phillips, J. C. *et al.* Scalable molecular dynamics with NAMD. *J Comput Chem* **26**, 1781-1802, doi:10.1002/jcc.20289 (2005).
241 Zhu, M. M., Rempel, D. L., Du, Z. & Gross, M. L. Quantification of protein-ligand interactions by mass spectrometry, titration, and H/D exchange: PLIMSTEX. *J Am Chem Soc* **125**, 5252-5253, doi:10.1021/ja029460d (2003).
242 Zhu, M. M., Rempel, D. L. & Gross, M. L. Modeling data from titration, amide H/D exchange, and mass spectrometry to obtain protein-ligand binding constants. *J Am Soc Mass Spectrom* **15**, 388-397, doi:10.1016/j.jasms.2003.11.007 (2004).
243 Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal Chem* **68**, 850-858 (1996).
244 Olsen, J. V. *et al.* Parts per million mass accuracy on an Orbitrap mass spectrometer via lock mass injection into a C-trap. *Mol Cell Proteomics* **4**, 2010-2021, doi:10.1074/mcp.T500030-MCP200 (2005).
245 Rand, K. D., Zehl, M., Jensen, O. N. & Jorgensen, T. J. Protein hydrogen exchange measured at single-residue resolution by electron transfer dissociation mass spectrometry. *Anal Chem* **81**, 5577-5584, doi:10.1021/ac9008447 (2009).
246 Wales, T. E. & Engen, J. R. Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass Spectrom Rev* **25**, 158-170, doi:10.1002/mas.20064 (2006).
247 Mistarz, U. H., Brown, J. M., Haselmann, K. F. & Rand, K. D. Probing the Binding Interfaces of Protein Complexes Using Gas-Phase H/D Exchange Mass Spectrometry. *Structure* **24**, 310-318, doi:10.1016/j.str.2015.11.013 (2016).
248 Marcsisin, S. R. & Engen, J. R. Hydrogen exchange mass spectrometry: what is it and what can it tell us? *Analytical and bioanalytical chemistry* **397**, 967-972, doi:10.1007/s00216-010-3556-4 (2010).
249 Zhang, Y., Rempel, D. L. and Gross, M. L. in *Hydrogen Exchange Mass Spectrometry of Proteins* (2016).
250 Masson, G. R., Jenkins, M. L. & Burke, J. E. An overview of hydrogen deuterium exchange mass spectrometry (HDX-MS) in drug discovery. *Expert Opin Drug Discov* **12**, 981-994, doi:10.1080/17460441.2017.1363734 (2017).
251 Quistgaard, E. M., Low, C., Moberg, P., Tresaugues, L. & Nordlund, P. Structural basis for substrate transport in the GLUT-homology family of monosaccharide transporters. *Nat Struct Mol Biol* **20**, 766-768, doi:10.1038/nsmb.2569 (2013).
252 Wisedchaisri, G., Park, M. S., Iadanza, M. G., Zheng, H. & Gonen, T. Proton-coupled sugar transport in the prototypical major facilitator superfamily protein XylE. *Nat Commun* **5**, 4521, doi:10.1038/ncomms5521 (2014).

Chapter 6: Appendix

6.1 Supplementary Information: Developing a workflow for modelling protein flexibility using ion-mobility MS and gas phase simulations

The original supplementary information files can be found online at:

https://onlinelibrary.wiley.com/doi/full/10.1002/anie.201812018

6.1.1 Supplementary Tables

Supplementary Table 6.1. Experimental values for IgG1 samples

Subclass	IgG from Sigma	Herceptin	Waters mAb standard
Theoretical mass (kDa)[h]	150	150	150
Experimental mass (Da)[i]	149,328 (±89)	148,620 (±64)	149,719 (±54)
Experimental charge[j]	21+	21+	20+
Glycosylated *CCS* (Å²)[k]	6827 (±81)	6875 (±50)	6883 (±57)
Deglycosylated *CCS* (Å²)[l]	6851 (±61)	6786 (±54)	6762 (±63)

[h] Approximate mass of glycosylated protein given glycoform variability, and sequence variability in Fc and Fab regions for Sigma IgG1.

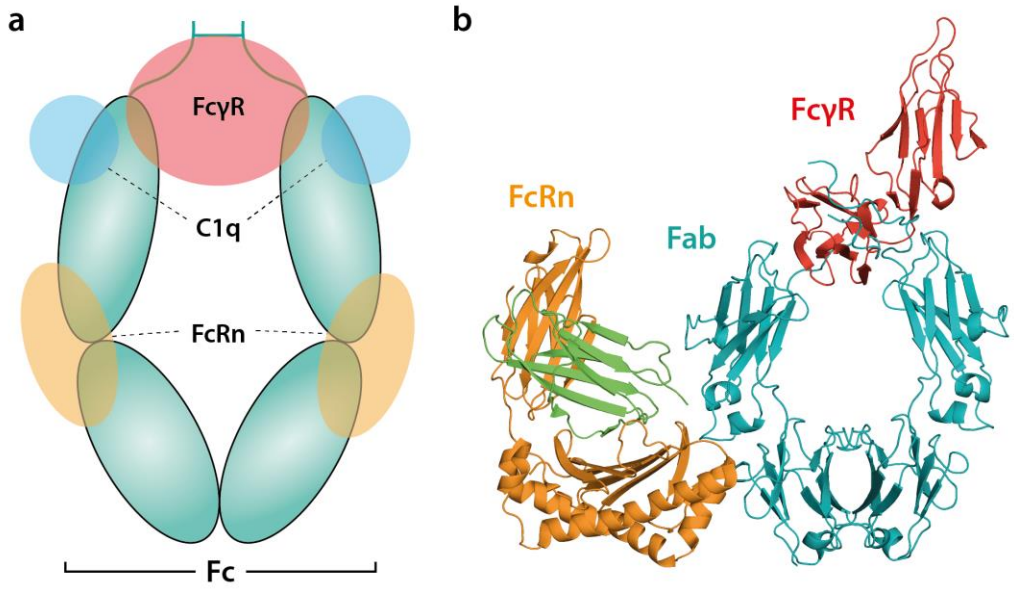
[i] Experimentally observed glycosylated mass via MS (± standard deviation).

[j] Lowest observed experimental charge.

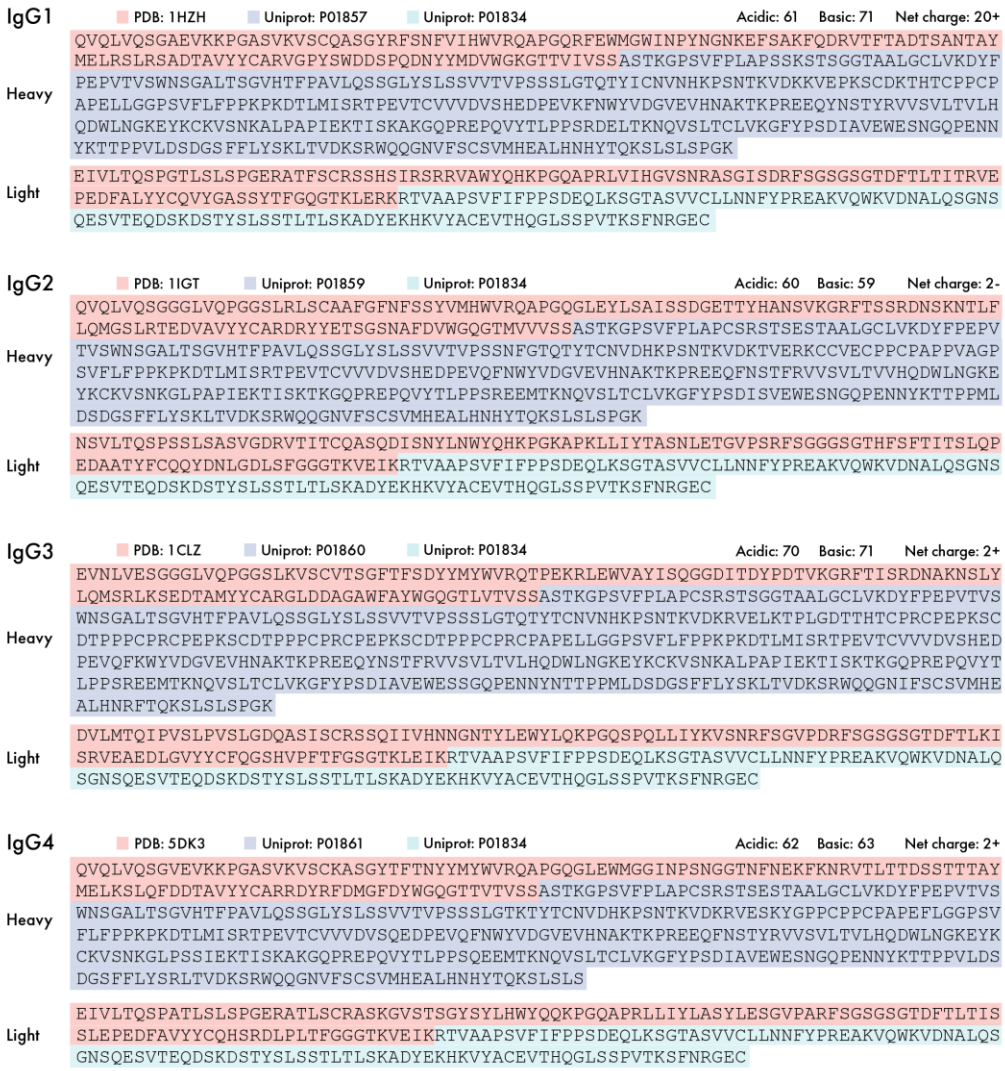
[k] Average CCS for lowest charge over T-waves 550, 600 and 640 ms^{-1} (± standard deviation) for glycosylated proteins.

[l] Average CCS for lowest charge over T-waves 550, 600, and 640 ms^{-1} (± standard deviation) for deglycosylated proteins.

6.1.2 Supplementary Figures

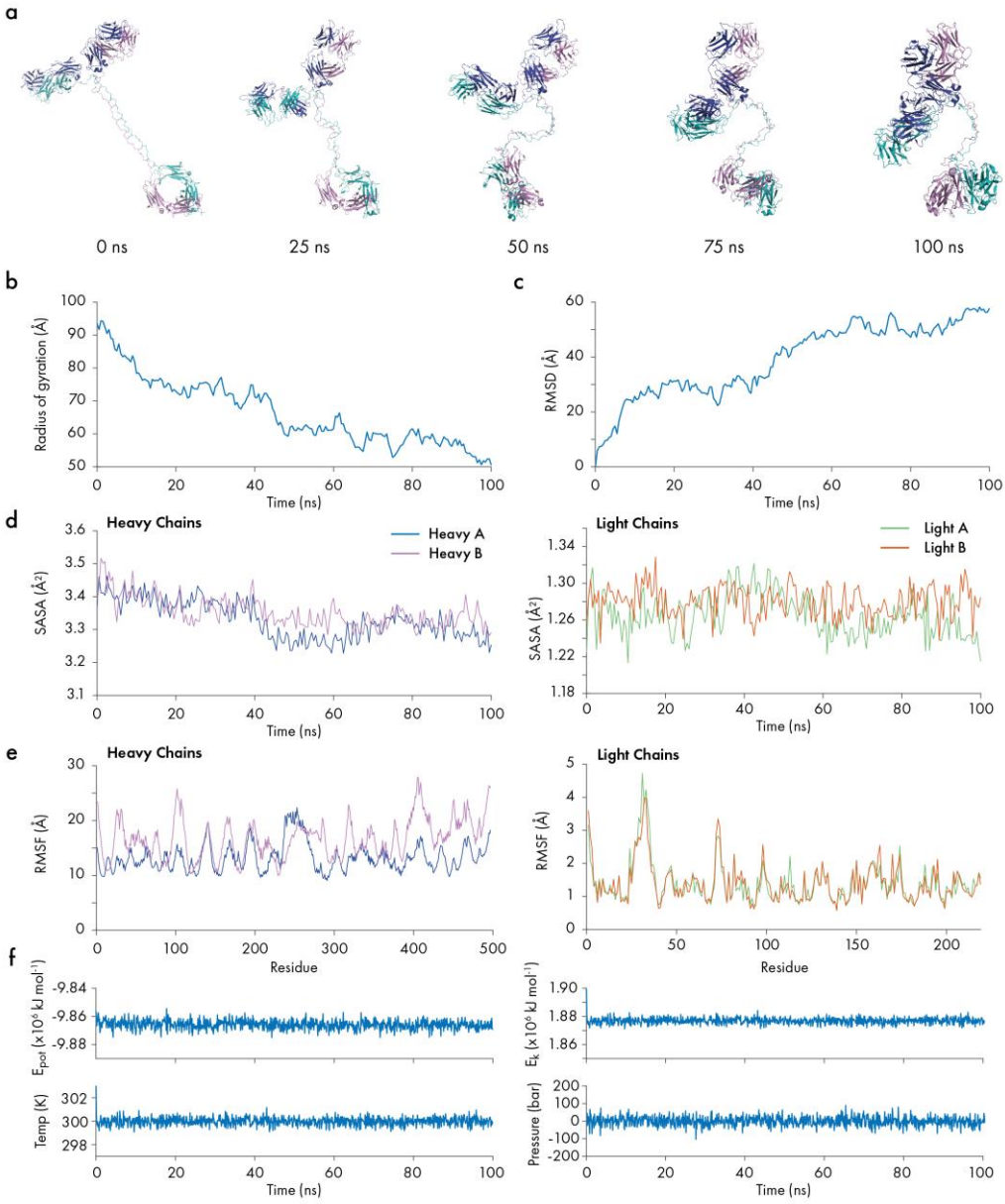


Supplementary Figure 6.1. Schematic of IgG Fc binding sites. (a) FcγR, C1q and FcRn binding sites are shown for the IgG Fc. (b) Alignment of FcγR crystal structure (orange/green; PDB 1T83) and FcRn (red; PDB 1FRT) to Fc of IgG2 (teal; PDB 1IGT). The hinge is located at the top of the Fc for both (a) and (b). C1q-Fc complex not shown due to the lack of crystallographic representation.

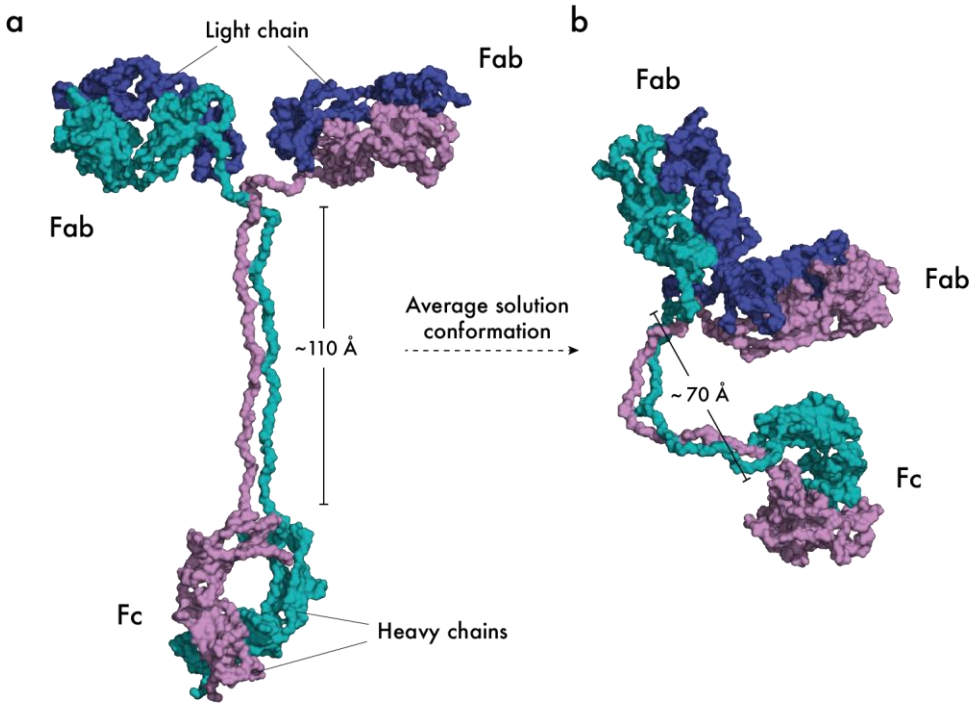


Supplementary Figure 6.2. Sequences of IgG1-4 homology models. The source of the variable and canonical sequences of each IgG heavy and light chain are shown. Variable regions (red) were inherited from crystal structures used for homology modelling for each IgG. Canonical sequences of the heavy (purple) and light (blue) chains were accessed from UniProt. The number of acidic and basic residues were calculated via the ProtParam [m] webserver for each half of the IgG molecule. The net charge displayed is calculated for the whole IgG molecule.

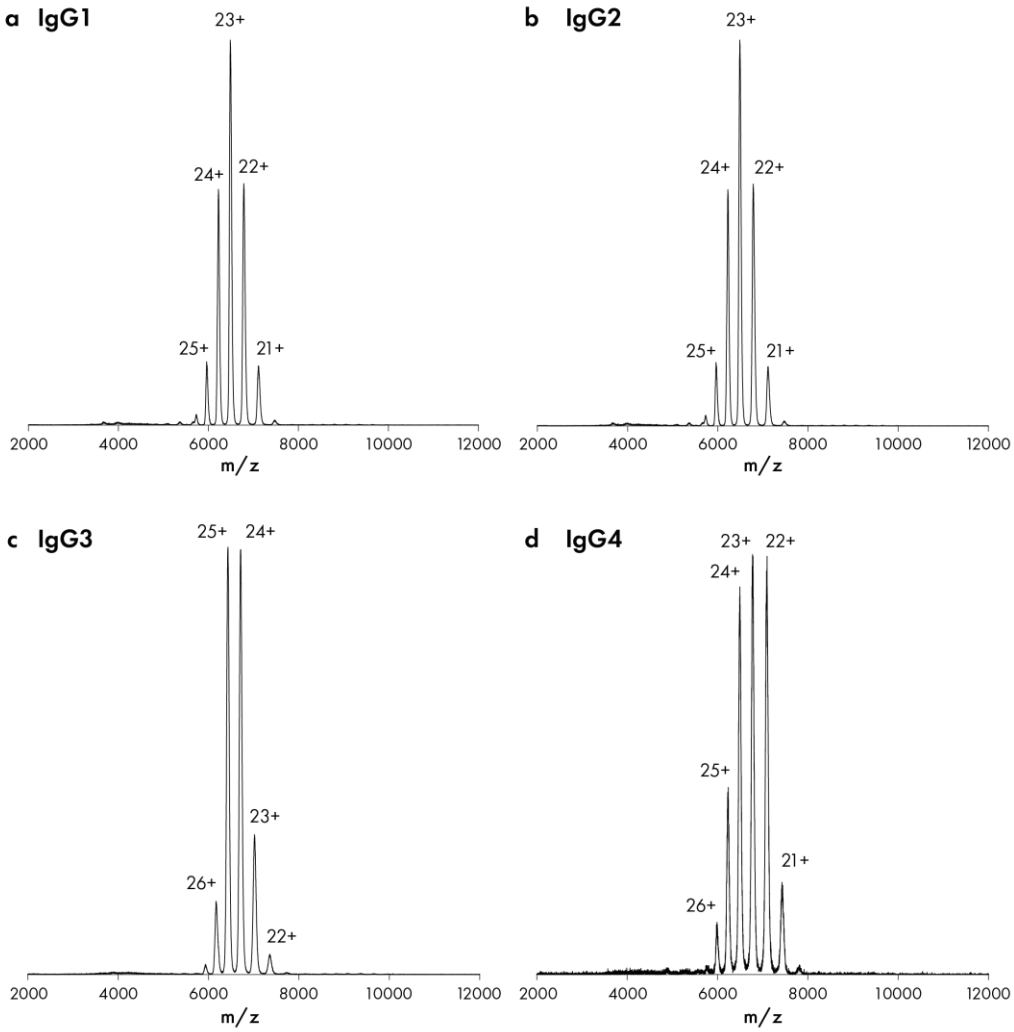
m https://web.expasy.org/protparam/



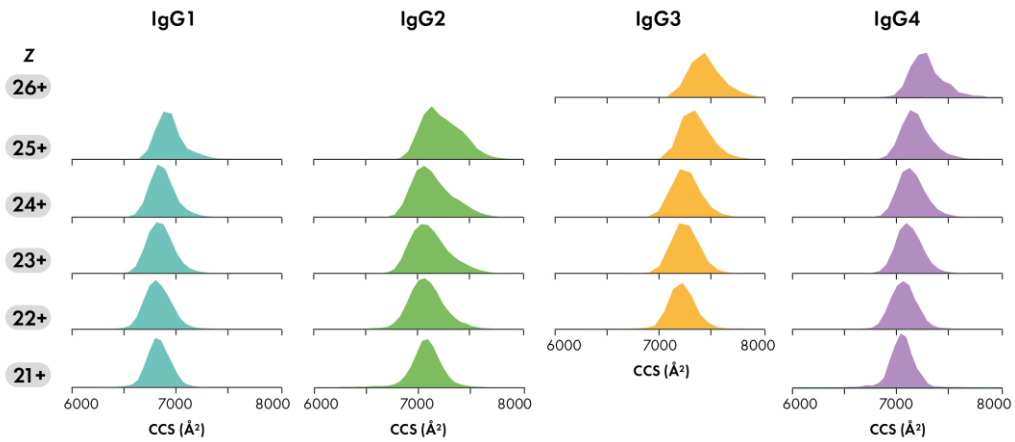
Supplementary Figure 6.3. Solution simulation of IgG3 homology model. (a) Timeline of IgG3 structure. Heavy chains are shown in teal and violet, light chains are in blue. Simulation parameters (b) radius of gyration, (c) RMSD, (d) solvent accessible surface area (SASA) and (e) root mean squared fluctuation of sidechain atoms over the simulation time. (f) System potential and kinetic energy, temperature and pressure parameters over simulation time. Analysis performed using GROMACS analysis tools[106].



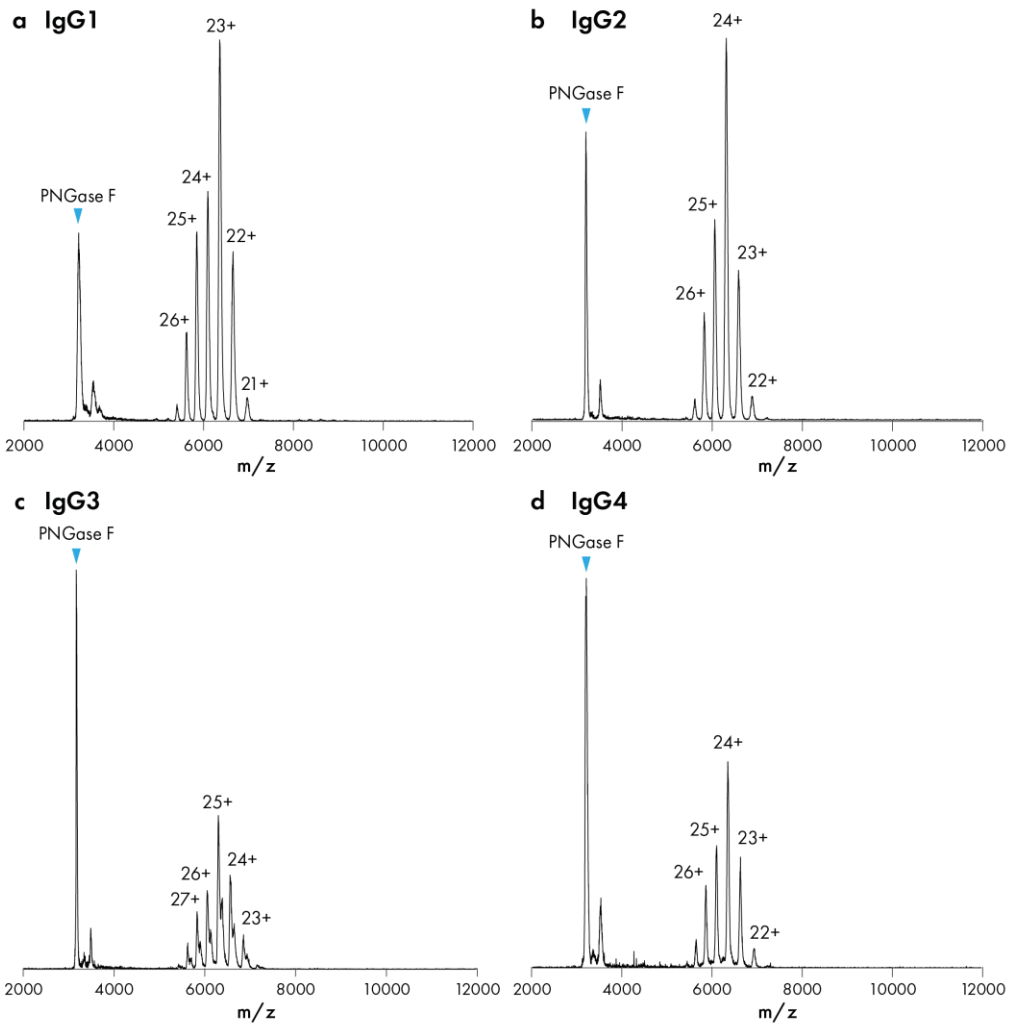
Supplementary Figure 6.4. Hinge length of human IgG3. (a) Homology model of human IgG3 showing the approximately 110 Å hinge region. (b) Average conformation of human IgG3 following 100 ns of solution simulation showing approximately 70 Å hinge region.



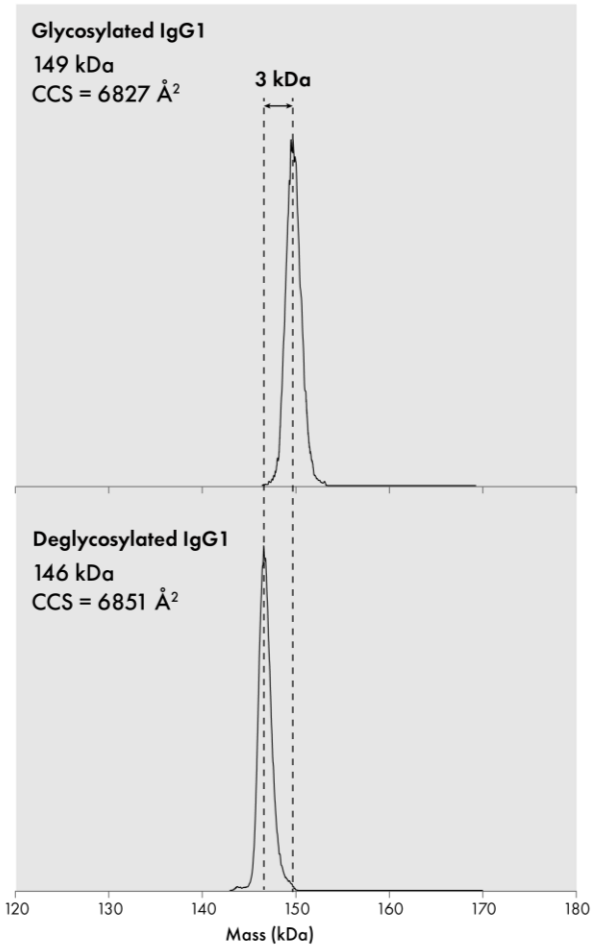
Supplementary Figure 6.5. MS spectra of glycosylated IgG1-4. Native mass spectra of all four subtypes of glycosylated IgG showing the charge envelope resulting from electrospray ionisation.



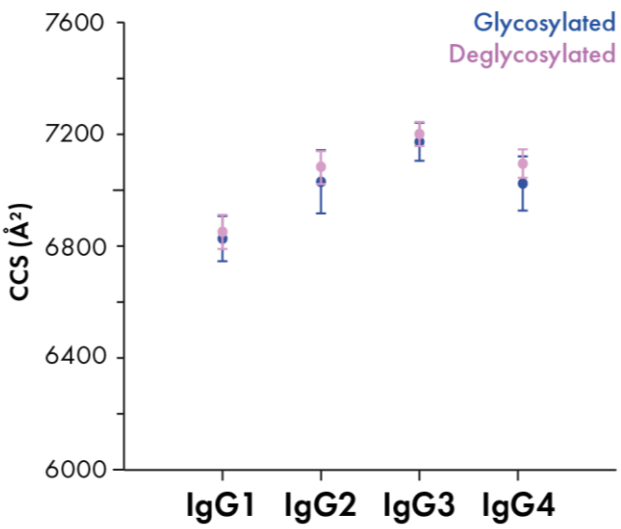
Supplementary Figure 6.6. *CCS* distributions of IgG1-4. Collisional cross section distributions for each charge state of IgG1-4 achieved using T-wave ion mobility. Higher charge states exhibit broader distributions.



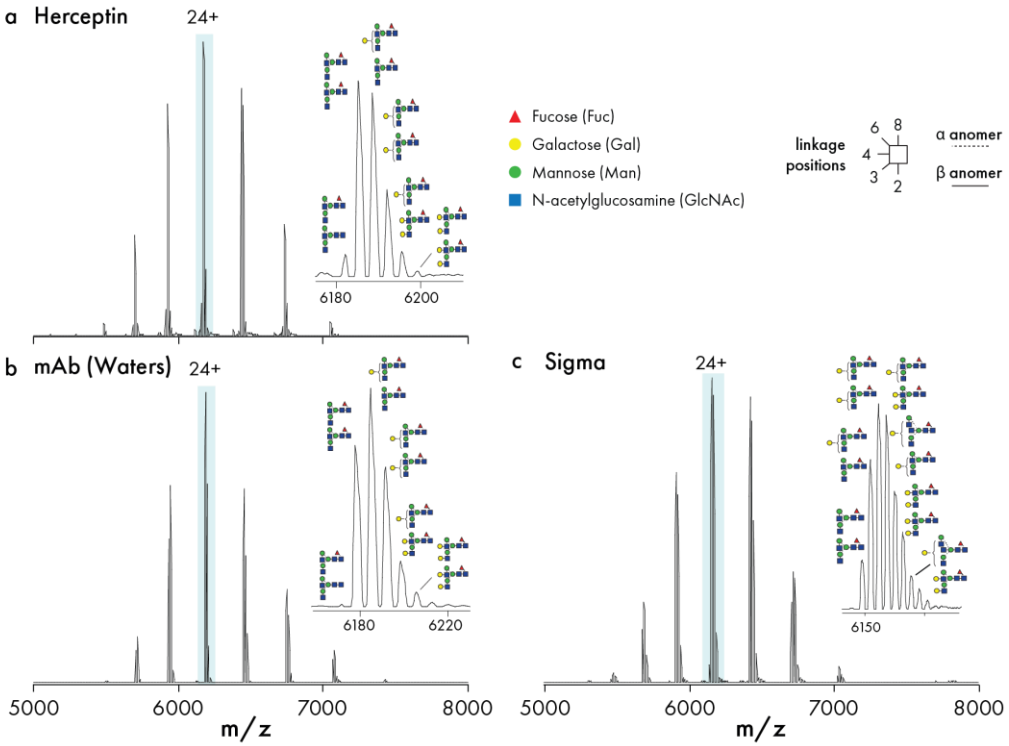
Supplementary Figure 6.7. MS spectra of deglycosylated IgG1-4. Native mass spectra of all four subtypes of deglycosylated IgG showing charge envelope resulting from electrospray ionisation. The deglycosylation enzyme (PNGase F) has been labelled at approximately *m/z* 3000.



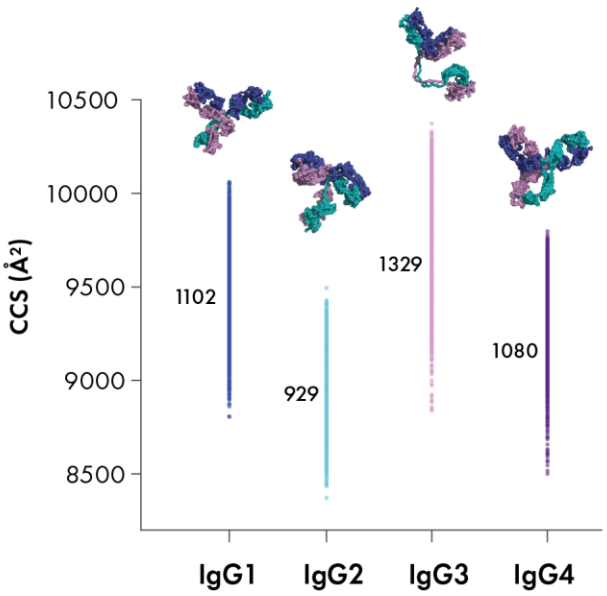
Supplementary Figure 6.8. Deconvoluted MS spectra of glycosylated and deglycosylated IgG1. Deconvolution performed using Waters MassLynx mass measurement tool and transform function. Mass shift of approximately 3 kDa is observed with no significant changes in *CCS*.



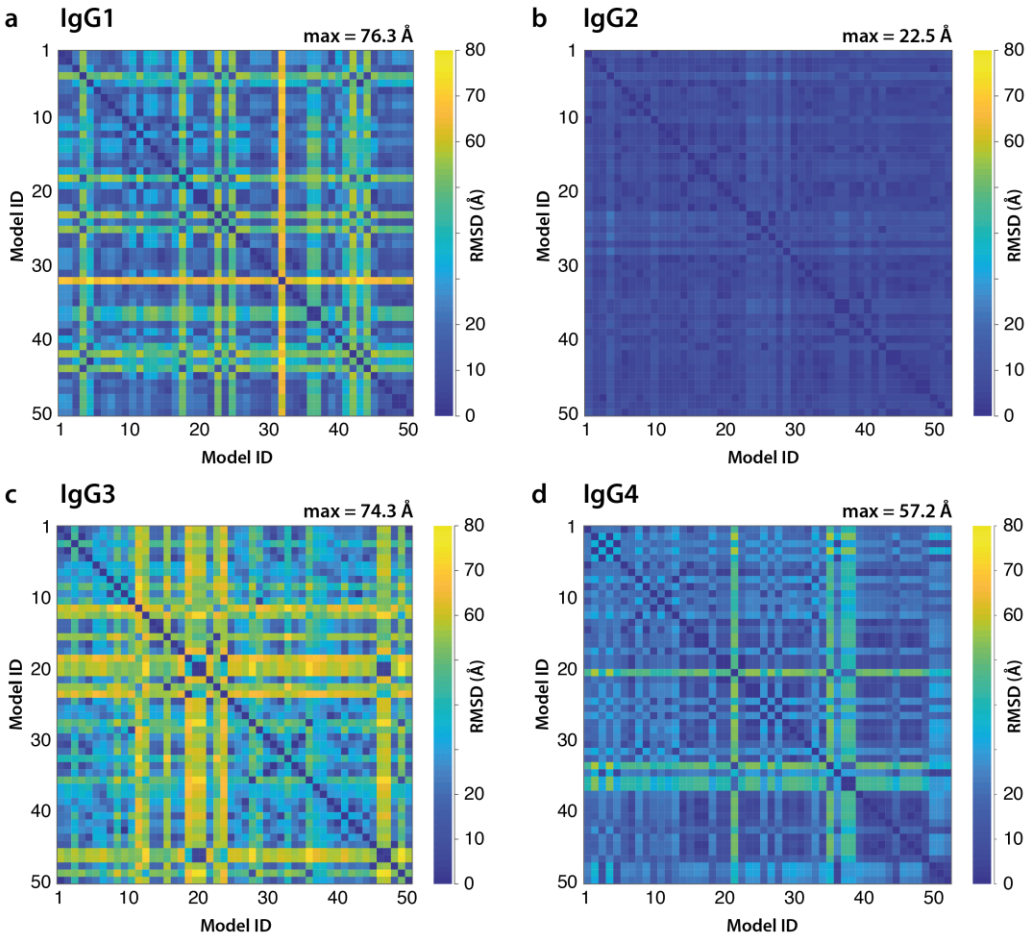
Supplementary Figure 6.9. CCS_{exp} of IgG1-4. CCS_{exp} for IgG1-4 shown for the lowest experimental charge states (21+ for IgG1, IgG2 and IgG4, 22+ for IgG3). Measurements were recorded and averaged for 550, 600 and 640 ms^{-1} T-wave velocities. Error bars show standard deviations of measurement across T-wave triplicates.



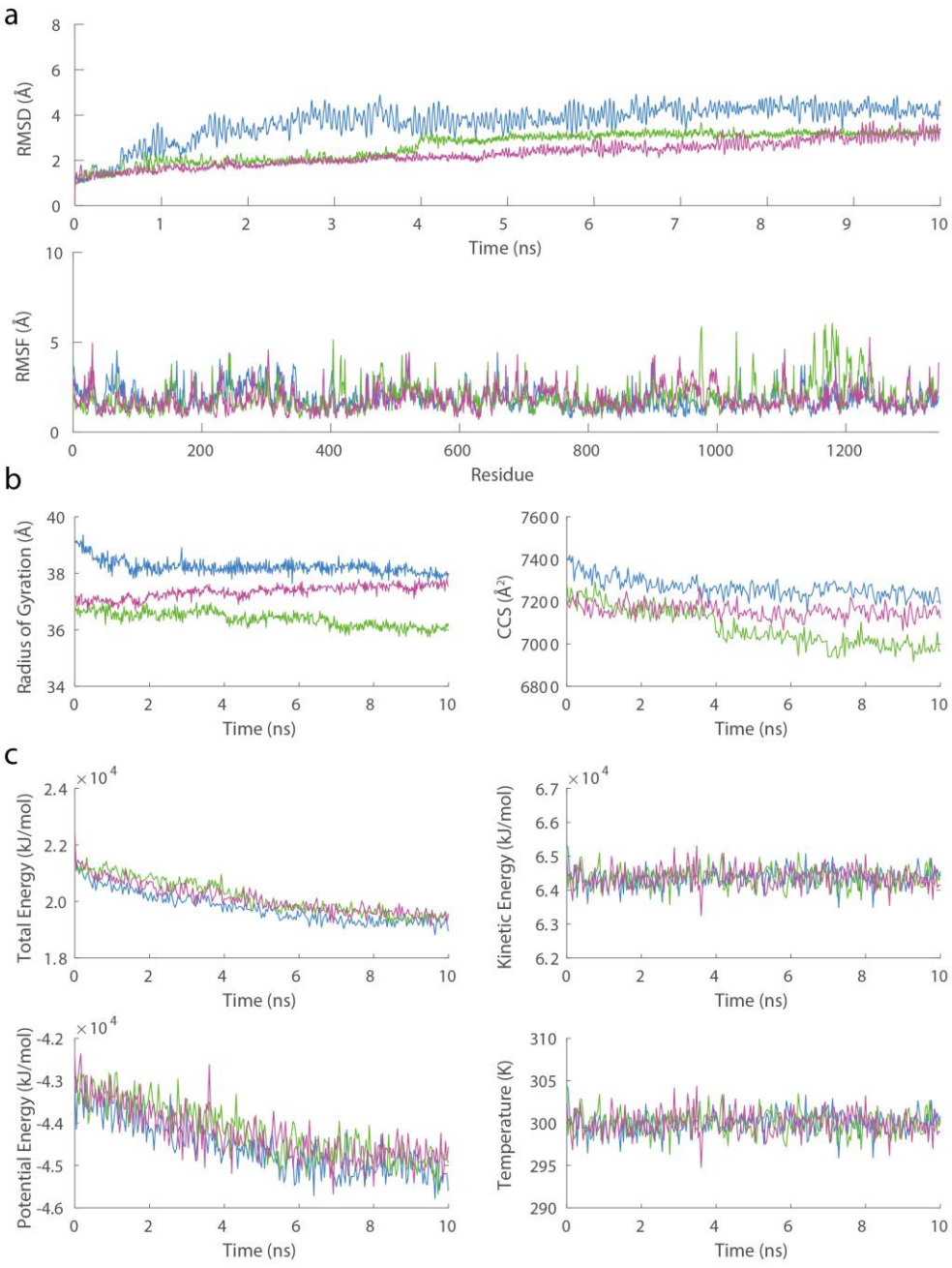
Supplementary Figure 6.10. High resolution native MS of Herceptin, Waters mAb and Sigma human plasma IgG1 samples, revealing glycoform heterogeneity. The individual glycoforms are displayed for the most abundant ion (inset). N-glycan structures have been labelled for each identified peak.



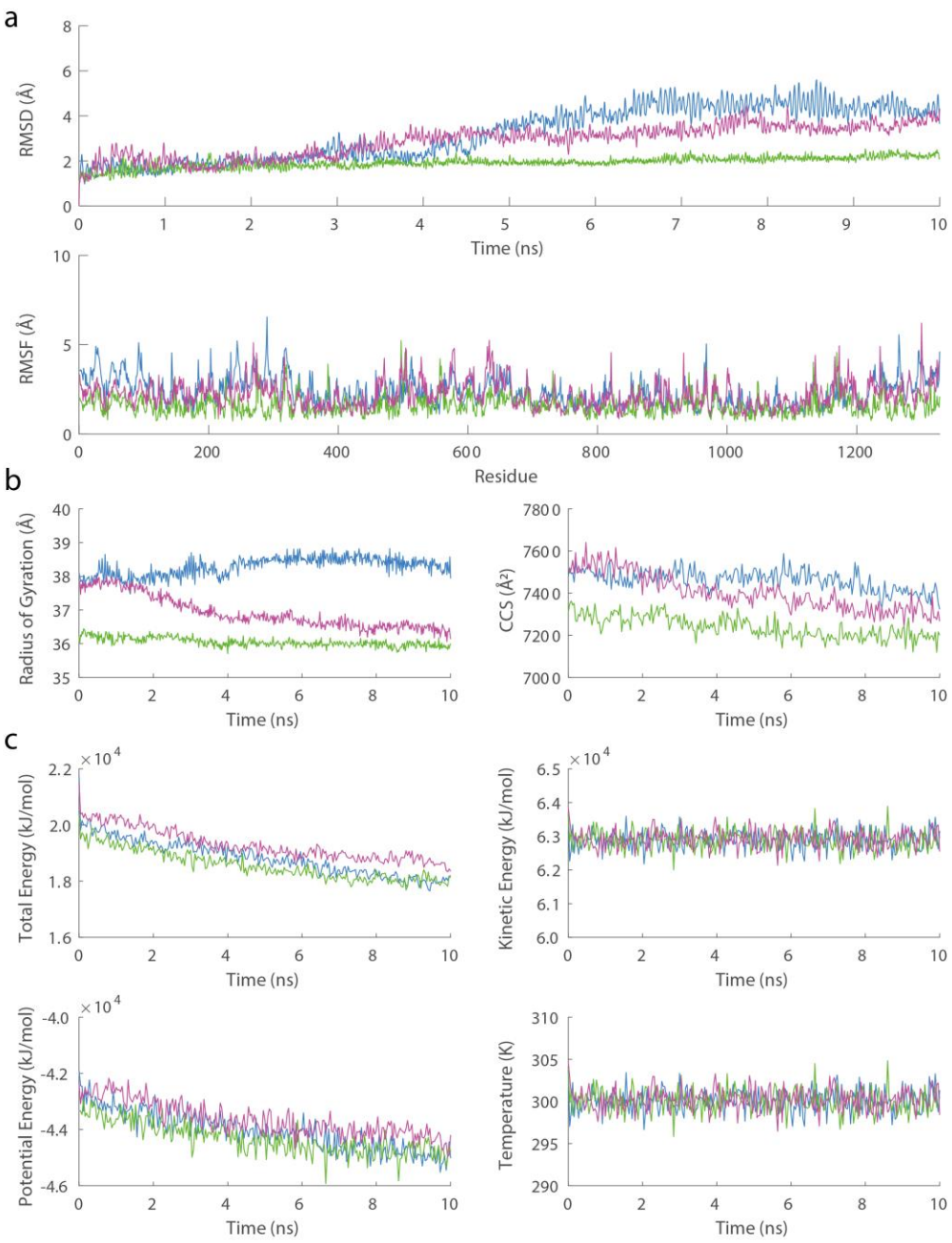
Supplementary Figure 6.11. *CCS* of all Fab conformations of IgG1-4 post-sampling. Values indicate the difference between largest and smallest *CCS* models (Δ*CCS*). All *CCS* calculated using IMPACT software[23].



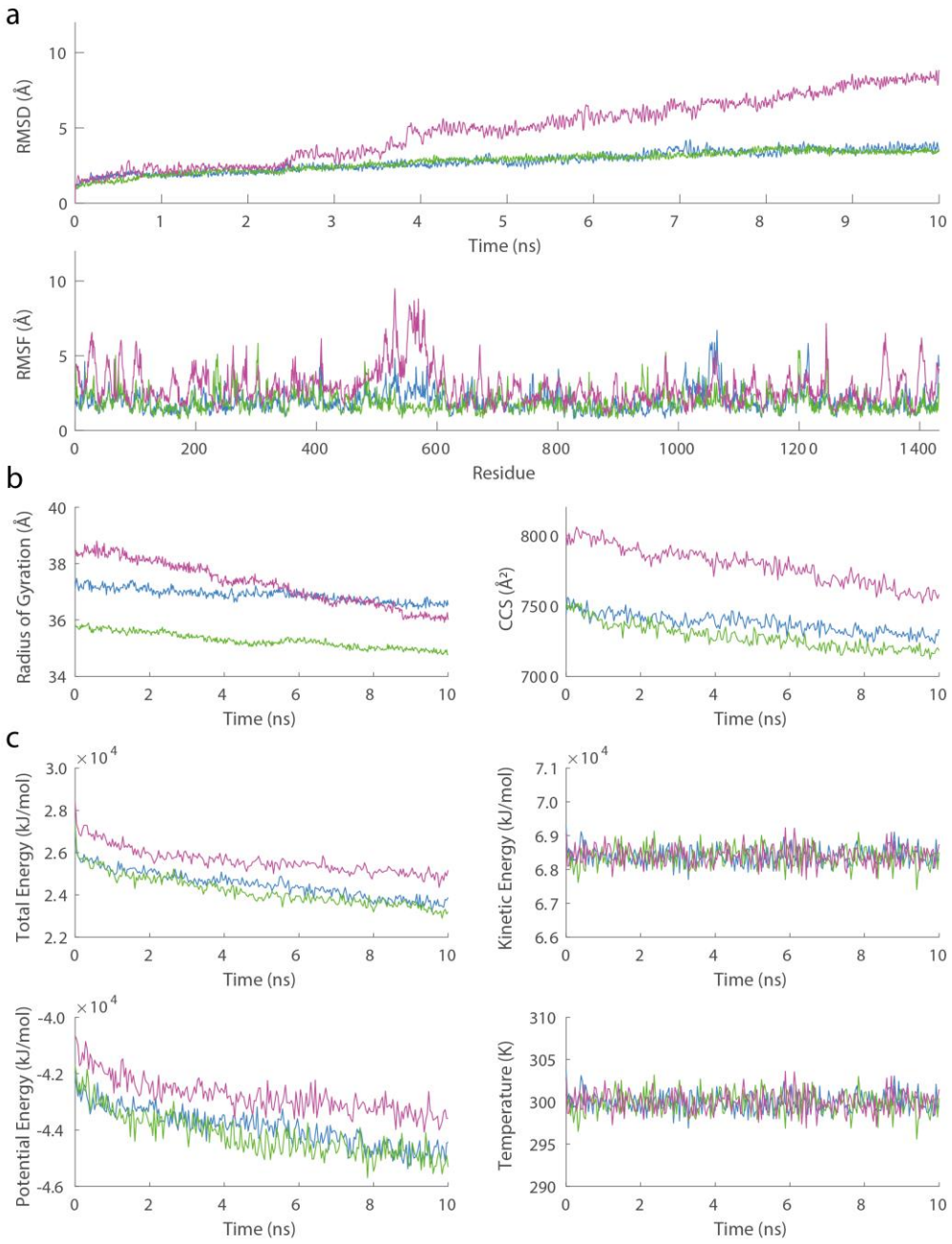
Supplementary Figure 6.12. RMSD matrices of 50 lowest *CCS* models of IgG1-4. Low RMSD variation between lowest *CCS* models of IgG2 compared to IgG1, IgG3 and IgG4 suggest a more confined conformational space populated by homogenous models. IgG3 model ensemble exhibits high RMSD variation frequently greater than 70 Å, indicating compact conformations can be varied. RMSD calculated between non-fitted Cα atoms only. Max RMSD is shown for each protein.



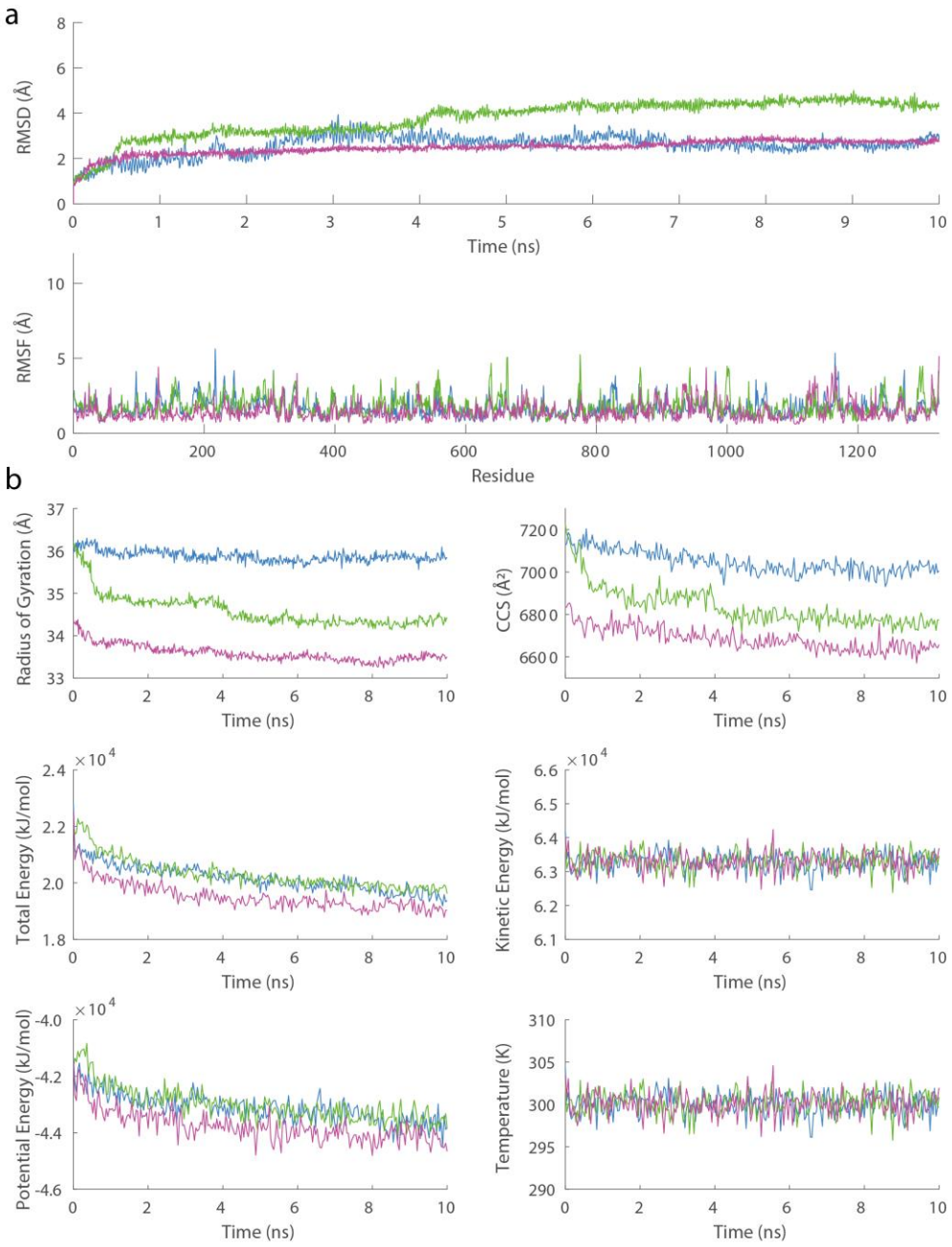
Supplementary Figure 6.13. Analysis of IgG1 gas phase simulations. (a) Structure similarity measurements RMSD and RMSF, (b) size measurements radius of gyration and *CCS*, and (c) system parameters over 10 ns simulation trajectories for each model.



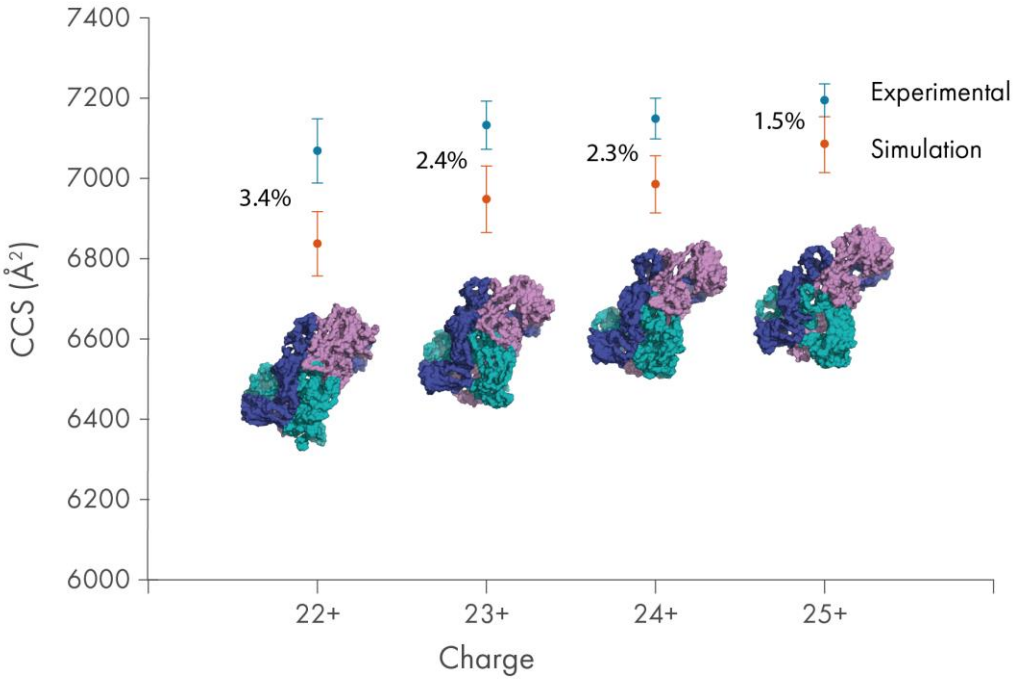
Supplementary Figure 6.14. Analysis of IgG2 gas phase simulations. (a) Structure similarity measurements RMSD and RMSF, (b) size measurements radius of gyration and *CCS*, and (c) system parameters over 10 ns simulation trajectories for each model.



Supplementary Figure 6.15. Analysis of IgG3 gas phase simulations. (a) Structure similarity measurements RMSD and RMSF, (b) size measurements radius of gyration and *CCS*, and (c) system parameters over 10 ns simulation trajectories for each model.



Supplementary Figure 6.16. Analysis of IgG4 gas phase simulations. (a) Structure similarity measurements RMSD and RMSF, (b) size measurements radius of gyration and *CCS*, and (c) system parameters over 10 ns simulation trajectories for each model.



Supplementary Figure 6.17. Gas phase simulations of IgG4 for charges 22-25+. Beginning from an identical IgG4 model, we performed four simulations, reflecting 22-25+ experimental charge states. Each model was pre-charged using a different random distribution of charge sites and simulated for 10 ns. The average *CCS* and *CCS* variation over the last 1 ns of each simulation is shown by the orange data point and error bars. The experimental *CCS* and standard deviation are shown in blue. The final model of each simulation is shown under each data point. The percentages represent the *CCS* percentage difference between the experimental and simulated *CCS* values.

6.2 Original publication: Deuteros: software for rapid analysis and visualization of data from differential hydrogen deuterium exchange-mass spectrometry

Andy M. C. Lau[1], Zainab Ahdash[1,#], Chloe Martens[1,#] and Argyris Politis[1,*]

[1]Department of Chemistry, King's College London, 7 Trinity Street, London, SE1 1DB, United Kingdom

*To whom correspondence should be addressed.

# These authors contributed equally.

6.2.1 Abstract

Summary: Hydrogen deuterium exchange-mass spectrometry (HDX-MS) has emerged as a powerful technique for interrogating the conformational dynamics of proteins and their complexes. Currently, analysis of HDX-MS data remains a laborious procedure, mainly due to the lack of streamlined software to process the large datasets. We present Deuteros which is a standalone software designed to be coupled with Waters DynamX HDX data analysis software, allowing the rapid analysis and visualization of data from differential HDX-MS.

Availability and implementation: Deuteros is open-source and can be downloaded from https://github.com/andymlau/Deuteros, under the Apache 2.0 license. Written in MATLAB and supported on both Windows and MacOS. Requires the MATLAB runtime library. According to the Wellcome Trust and UK research councils' Common Principles on Data Policy on data, software and materials manage- ment

and sharing, all data supporting this study will be openly available from the software repository.

6.2.2 Introduction

Hydrogen deuterium exchange mass spectrometry (HDX-MS) is a structural technique which has garnered attention for its ability to assess protein-protein and protein-ligand interactions, protein folding, and the associated dynamics of these processes[64,65,247,250]. The basis of HDX-MS relies on the exchange of labile amide hydrogens of the protein backbones, for bulk deuterium within solution. The protein of interest is allowed to undergo exchange in a deuterium-rich buffer for a set number of timepoints and then quenched. The protein is then enzymatically cleaved to the peptide level, and the mixture is subjected to liquid chromatography coupled to mass spectrometry (LC-MS). Using LC-MS, the mass of the peptide acquired through deuteration can be determined via a database search. Peptides which participate in hydrogen bonding of amide hydrogens result in lesser exchange[248]. Additionally, those which comprise the accessible surfaces of proteins, may experience relatively greater deuteration, than those found in the protein interior[248]. In differential HDX-MS (ΔHDX-MS), peptides from a reference state are compared with those from an altered state (which could be for instance a mutation or a ligand) to report on regions of the protein which are affected by structural or conformational perturbations.

After data acquisition, the analysis of HDX-MS data using Waters Instrumentation consists of several steps: (i) peptide identification using the Protein Lynx Global Server (Waters Corp.), (ii) peptide mass assignment using DynamX (Waters Corp.) and (iii) visualization and interpretation of data. In particular, interpretation of the results can be challenging due to the size of the datasets involved. A typical HDX-MS experiment will result in the order of 10^2 peptides depending on the system size and complexity. Several visualization methods have been introduced to provide clarity

on the ensemble of peptides generated from HDX-MS. Waters offers in DynamX, 'butterfly' and 'difference' plots which while useful during data analysis, have several shortcomings. They do not show the identity of each peptide, their lengths or regional peptide redundancy. As an alternative, the 'Woods' plot[126] (developed separately from Waters), provides a per-timepoint breakdown of the peptide ensemble, with each subplot displaying peptide length, start and end residues, global coverage and a vertical axis metric which may be absolute uptake (in Daltons) or relative fractional uptake (RFU). HDX data can additionally be visualized on uptake maps, and molecular representations where uptake and other data are projected onto 3-dimensional structures of the system[131].

To simplify the HDX-MS analysis workflow, we have developed Deuteros for the rapid visualization of differential HDX-MS data. Development of the software was primarily motivated by the lack of data representation methods for differential HDX-MS. Deuteros has been designed to be used post-DynamX, for the downstream statistical filtering and visualization of data from differential HDX-MS. It provides alternative visualization capabilities in the form of statistically filtered 'Woods' plots and through PyMOL-compatible scripts, which allow differential HDX-MS data to be projected onto three- dimensional structures of proteins. A combination of these visualization methods provides users with the ability to effortlessly identify biologically interestingly regions of proteins. Finally, inputs to Deuteros have been standardized to csv files, allowing the software to be potentially compatible with any instrumentation.

We have applied our software to a comparison of the wild-type xylose transporter (XylE) and a E153Q mutant. XylE is a secondary membrane transporter protein tasked with the role of shuttling xylose sugar across bacterial cell membranes[251]. A member of the Major Facilitator Superfamily (MFS), XylE operates through an alternating-access mechanism, transitioning be- tween inward-facing and outward-facing conformational states in a highly dynamic fashion[252].

6.2.3 How does it work?

Deuteros is a standalone MATLAB application available to both MacOS and Windows and requires the MATLAB runtime library. Deuteros is designed to assist interpretation of HDX-MS data that has been analyzed using DynamX (Waters Corp.) and provides complementary data visualization capabilities. Deuteros requires two inputs: the 'state' and 'difference' files exported from DynamX. Deuteros consists of four steps including data input and three visualization stages: flattened data maps, Woods plot (with statistical peptide filtering) and output to PyMOL.

6.2.4 Input data

The DynamX 'state' file contains a per-protein, per- peptide, per-time point aggregation of peptide deuterium uptake data from the ΔHDX-MS conditions. State files contain information including *m/z*, maximum possible deuterium uptake, observed deuterium uptake, standard deviation, retention time, and any residue modifications reported. Users should only enable proteins and states of interest and disable all others within the DynamX session file. The 'difference' file contains a per-peptide, per-timepoint comparison of peptide deuterium uptake from two user defined states. The difference file can only be generated from DynamX when two or more states are loaded into the dataset. Users should ensure that the correct comparison is made by selecting the correct states within DynamX. A video tutorial and example datasets have been provided alongside the software.

6.2.5 Visualization

Deuteros produces flattened data maps including coverage, residue-level redundancy, deuterium uptake heat maps and Woods plots. Coverage, redundancy and various deuterium uptake styles can also be exported from Deuteros, to be projected onto atomic models of the protein of interest in the PyMOL molecular graphics viewer[63]. Users can simply 'drag and drop' these files into PyMOL (for

MacOS, or PyMOL version 2.0 and above for Windows), or alternatively copy and paste the con- tents of the file into the PyMOL command line.

6.2.6 Statistics

Confidence limits are calculated as[140]:

$$CI_t = 0 \pm \left(\frac{\sigma_t}{\sqrt{N}}\right)\alpha \tag{6.1}$$

where σ_t is the standard deviation of the mean uptake for timepoint t, N is the number of sample replicates, and α is the critical value desired. By default, Deuteros provides critical values for 98 and 99% confidence limits (6.965 and 9.925) for a two-tailed t-test with 2 degrees of freedom. For 'sum' data, where peptide deuterium uptake differences from each timepoint are aggregated together to better identify potential peptides that are conformationally active, the confidence limits are calculated using:

$$CI_{sum} = 0 \pm \left(\frac{\sqrt{\sum_{t=1}^{n}\sigma_t^2}}{\sqrt{N_n}}\right)\alpha \tag{6.2}$$

where σ_t^2 is the variance of all peptides for timepoint t, N is the number of timepoint observations for the variance, n is the number of timepoints and α is the critical value as in (6.1). Errors are propagated as a simple sum of variables as according to Houde *et al*.[140].

6.2.7 Application

To showcase the capabilities of Deuteros, we imported state and difference files from the wild-type and E153Q mutant XylE membrane transporter Martens *et al*.[141]. The coverage map of XylE indicate a 92.5% coverage, with the largest non-covered region around residues 230–240 (Figure 6.1a). The redundancy map expands on the coverage map, displaying the same coverage, but with a white-magenta color gradient to represent peptide redundancy. Reviewing the map shows that the

highest redundancy of the XylE dataset was at 12 peptide copies around residues 465 (Figure 6.1b). The N-terminal, residues 90, 180, 260 and 360 have only a single peptide representing these regions.

The Woods plot section displays a per-timepoint breakdown of the differential dataset in a grid layout. Deuteros can display a maximum of approximately 8 timepoints simultaneously before individual Woods plots become too crowded, depending on the screen size and resolution. Woods plots first apply confidence filtering to all peptides in each timepoint (Figure 6.1c). Peptides with differential deuteration outside of the user selected confidence limits are non-significant and are shown in grey. The significant peptides are shown as red for deprotected and blue for those that are protected. While only one set of confidence limits are applied to the data, two boundaries are shown on each Woods plot as a visual aid for users to view which peptides might be significant, should they wish to tighten or relax the filter used. The legend section displays the confidence limits as 6 Da (to two decimal places) values around 0 (or no difference). To facilitate interpretation, significant peptides can also be exported as a csv file containing a per-peptide per-timepoint breakdown of the ΔHDX-MS data. Users may also take advantage of the in-built MATLAB data cursor which displays the residue number and differential uptake of a peptide by clicking on the peptide within the graphical user interface.

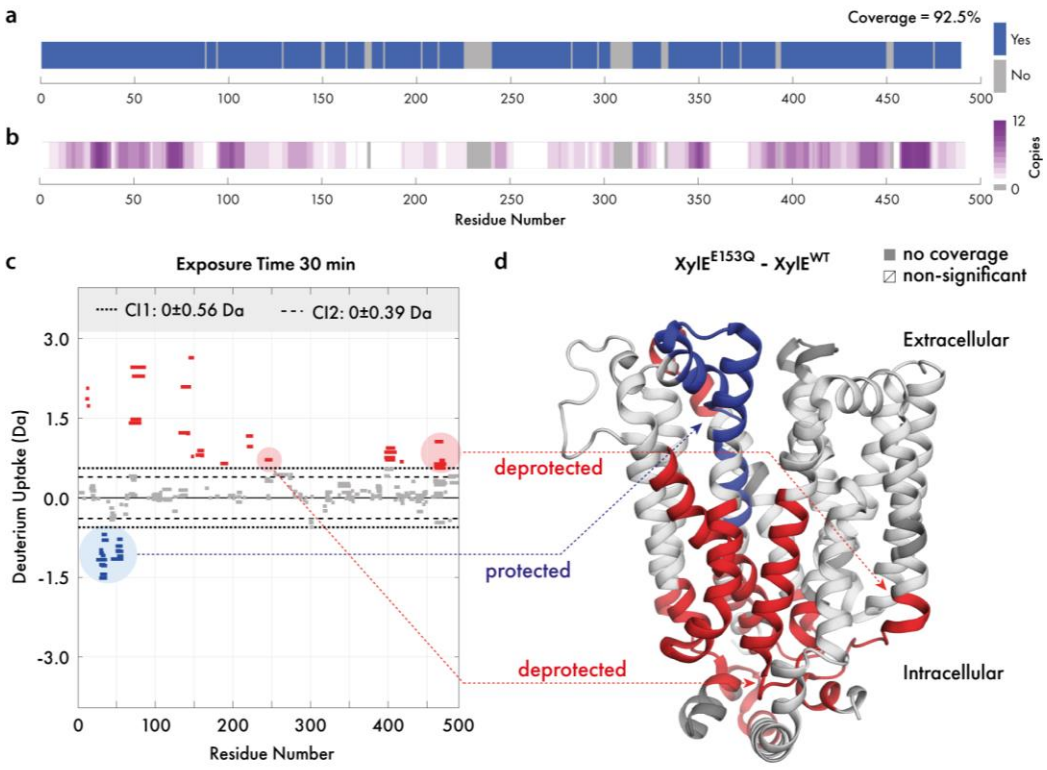


Figure 6.1. Overview of Deuteros demonstrated on the XylE transporter. Visualization of (a) experimental protein coverage, (b) data redundancy and (c) deuterium up- take differences in Woods plot format. Dashed and dotted lines indicate 98 and 99% confidence limits applied to the dataset to identify peptides with significant deuteration differences. Deprotected, protected and non-significantly different peptides are in red, blue and grey respectively. (d) Differential HDX-MS data for the wild-type and E153Q mutant XylE has been projected onto its crystal structure (PDB ID: 4GBY)

The PyMOL export section consists of options for formatting the data from the linear coverage map and Woods plot sections, for visualization in PyMOL through pml files. Coverage and redundancy can be projected onto structures and a range of color palettes are available. Differential deuteration data can also be exported for projection onto the molecular structure of XylE (PDB ID: 4GBY; Figure 6.1d). For this representation, the deuteration data type can show absolute differential uptake (in Daltons), or the differential relative fractional uptake (ΔRFU). The ΔRFU considers the peptide length and its maximum deuteration and scales the absolute uptake as a percentage of this value, which may be more informative for some datasets. Similar to the Woods plot, Deuteros implements red/blue/white/grey color scheme for

protected, deprotected, non-significant and non-covered regions. Through projection of deuteration data onto the structure of XylE, structural effects caused by the E153Q mutation are immediately visible (Figure 6.1d). The extracellular-facing portion of XylE experiences protection (blue), while the intracellular portion experiences deprotection (red).

Acknowledgements

We would like to acknowledge Jurgen Claesen (Hasselt University) for critical review of the manuscript and Eamonn Reading (King's College London) for suggestions with software features.

Funding

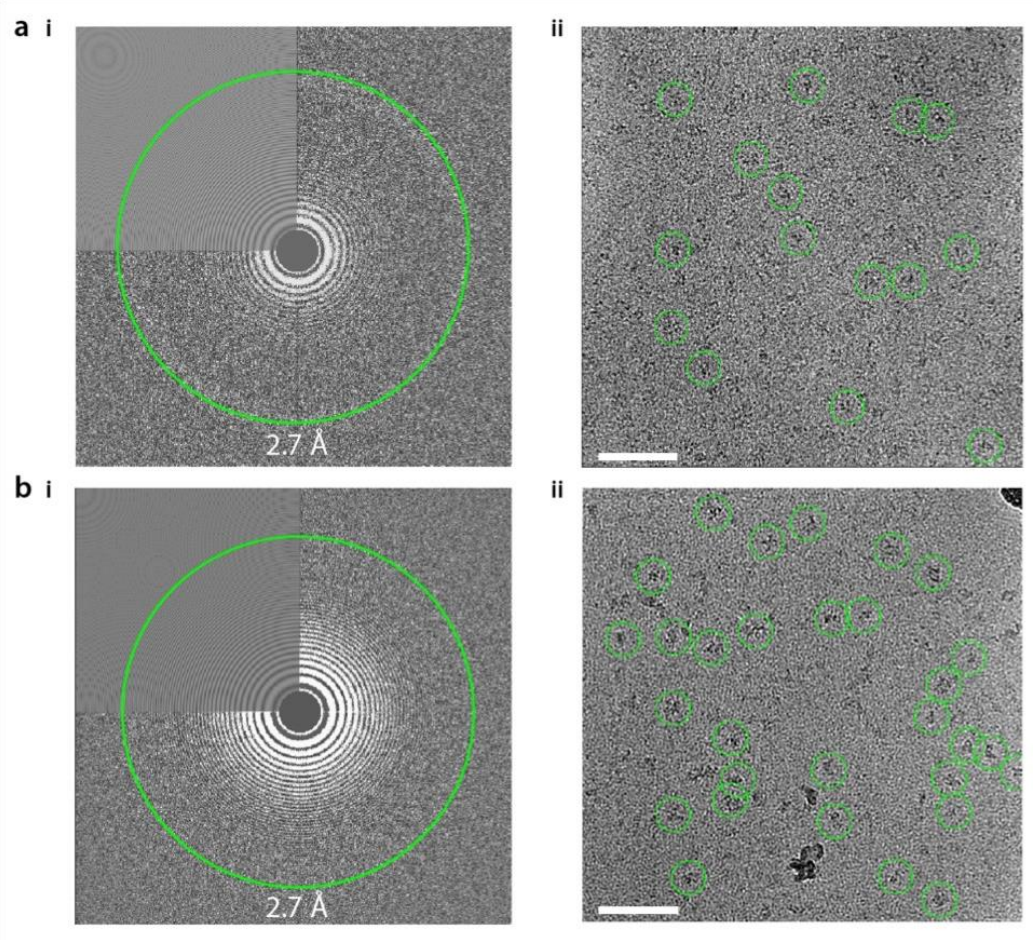
This work was supported by grants from the London Interdisciplinary Biosciences Consortium (LIDo) BBSRC Doctoral Training Partnership (BB/ M009513/1), the Wellcome Trust (109854/Z/15/Z) and the Medical Research Council (MC_PC_15031).

6.3 Supplementary Information for *Structural basis of the Cullin 2 RING E3 ligase regulation by the COP9 signalosome*

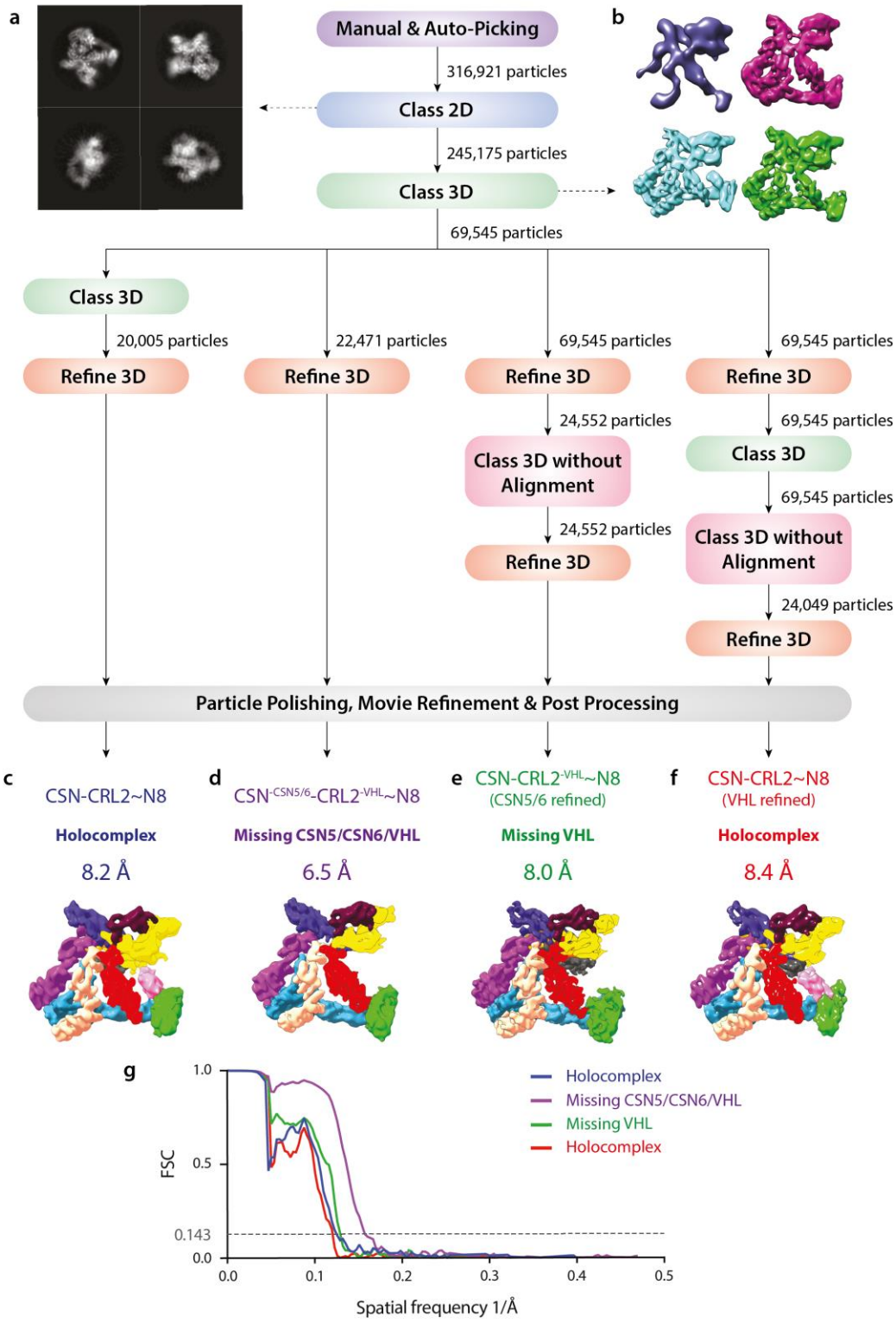
The original supplementary information files can be found online at:

https://www.nature.com/articles/s41467-019-11772-y#Sec26

6.3.1 Supplementary Figures

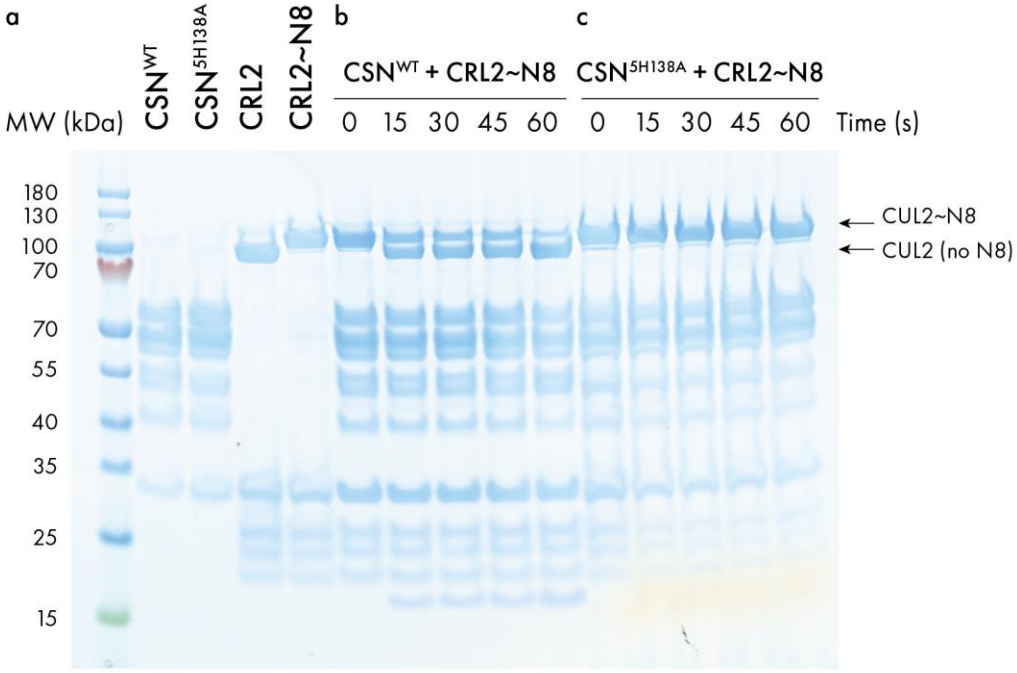


Supplementary Figure 6.18. Cryo-electron micrographs. Micrographs of (a) CSN-CRL2~N8 and (b) CSN-CRL2 complexes. (i) Representative power spectrum and (ii) various single molecular views are circled. Scale bar represents 80 nm.

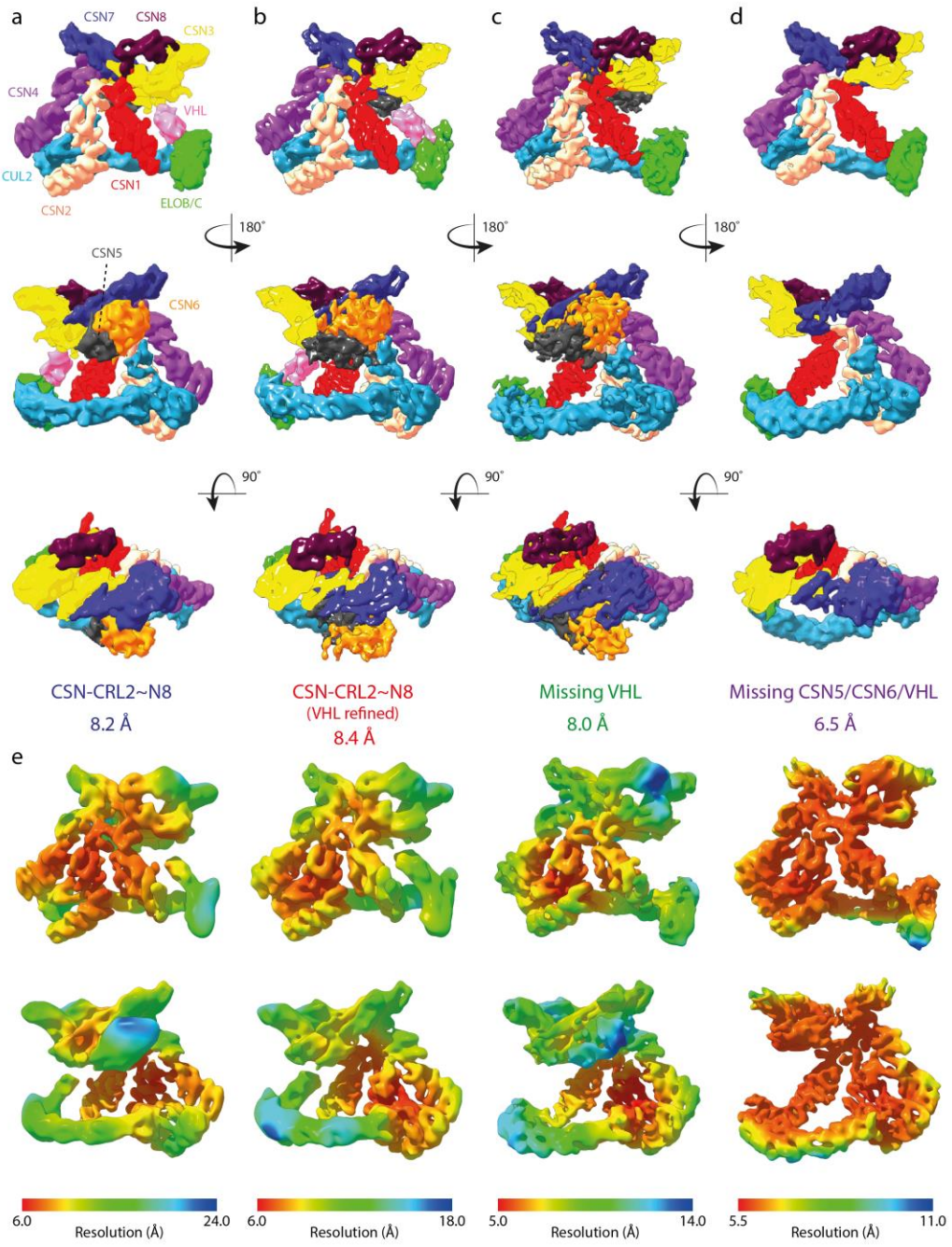


Supplementary Figure 6.19. Flowchart depicting the workflow for processing cryo-EM data. A set of ~3100 micrographs were subjected to manual and auto-picking in order to acquire particles for 2D reference-free classification (a). 2D classification was used for the positive selection of particles prior to 3D classification. (b) Four of the fifteen classes generated, demonstrate subunit

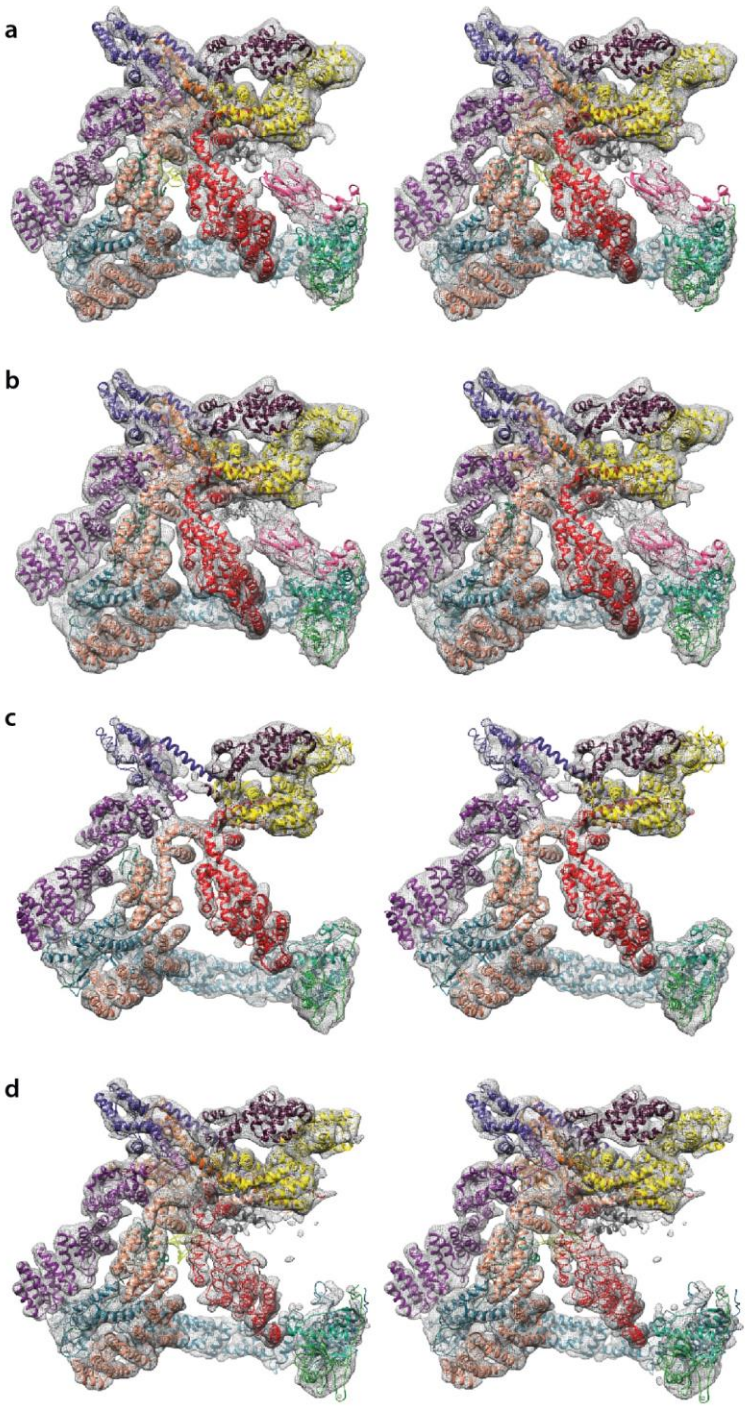
heterogeneity and presence of a small component of apo-CSN (purple model) in the data set. Particles from the three models containing the CSN and CRL2~N8 in (b) were pooled and further processed as described in the workflow. Maps (e) (CSN-CRL2^{-VHL}~N8) and (f) (CSN-CRL2~N8) were generated using focused refinement by applying a mask and using the area of interest as a starting model. (g) shows the Fourier shell correlation (FSC) for the four maps (c-f).



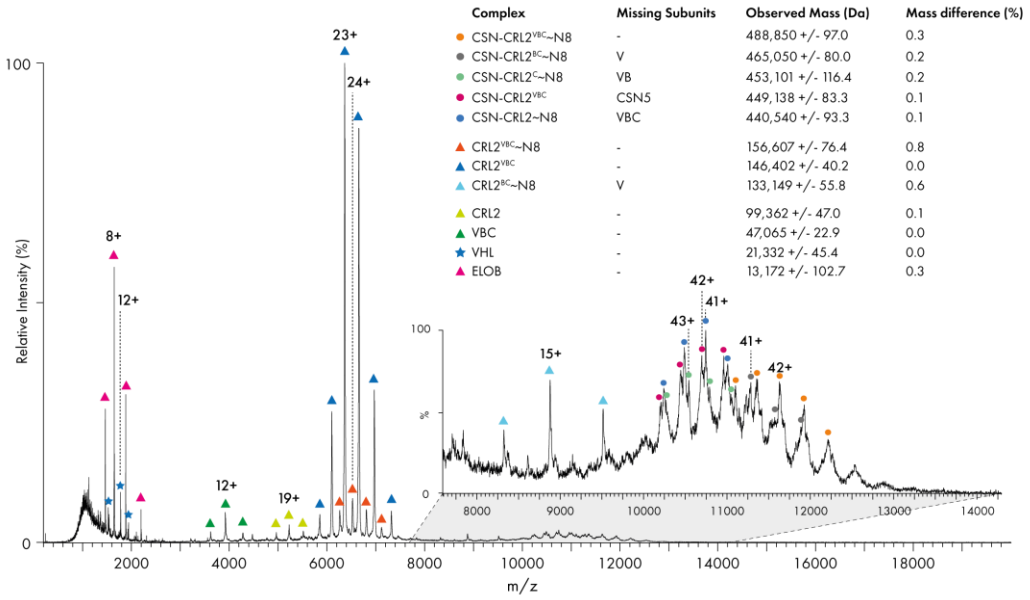
Supplementary Figure 6.20. Deneddylation activity of CSN^{WT} and CSN^{5H138A}. (a) Bands corresponding to denatured CSN and CRL2 complexes. (b) Incubation of CSN^{WT} with CRL2~N8 and (c) CSN^{5H138A} with CRL2~N8 over time. Proteins were incubated at 37˚C and reactions were inhibited through the addition of lithium dodecyl sulphate (LDS) and quickly heating to 90˚C using a pre-heated heat block to ensure rapid denaturation. Bands corresponding to CUL2~N8 and CUL2 are indicated by arrows for clarity.



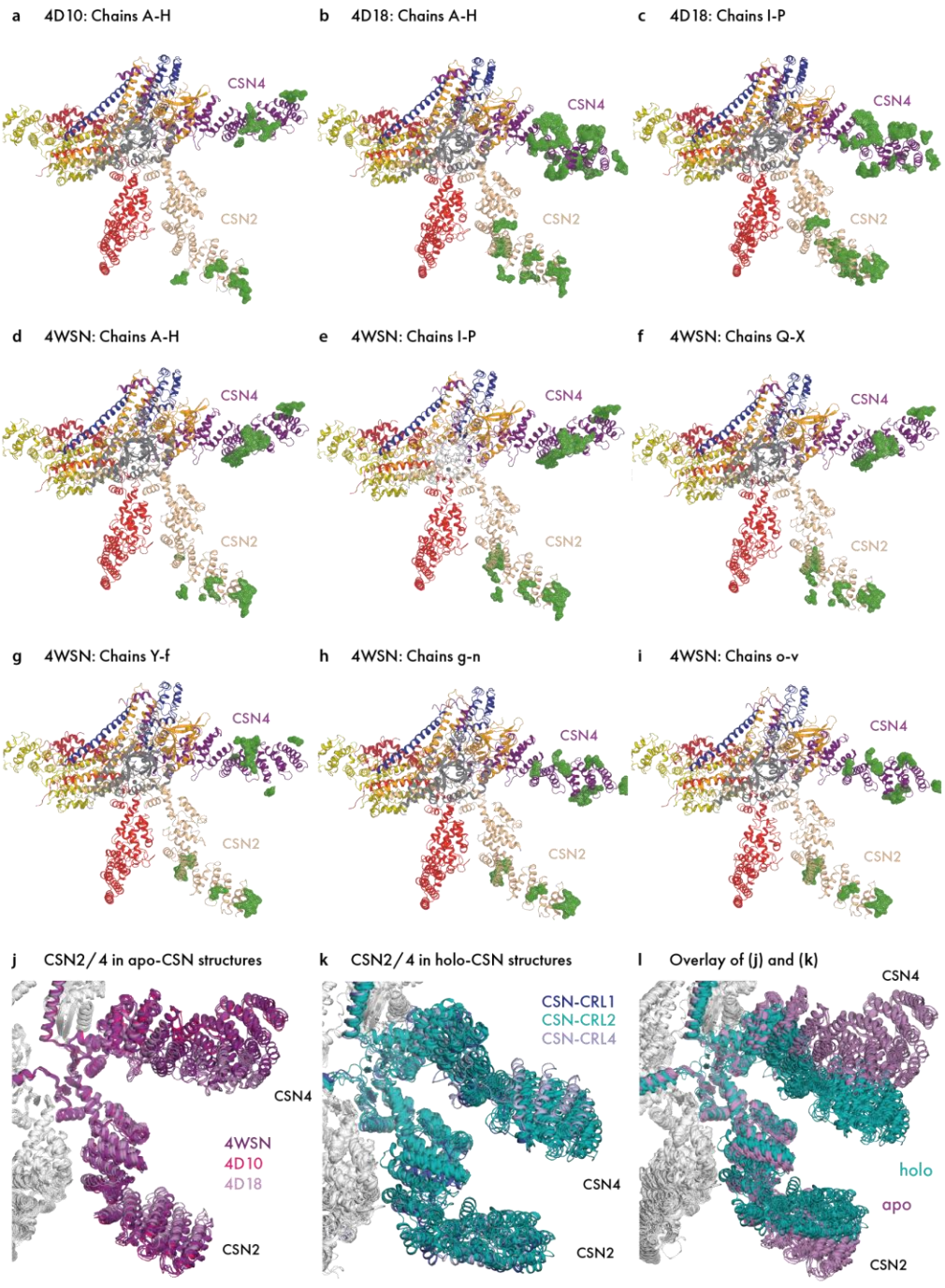
Supplementary Figure 6.21. Cryo-EM structures of the CSN-CRL2~N8 segmented to highlight subunit composition. Maps of the (a) CSN-CRL2~N8 holocomplex, (b) holocomplex (from VHL focused refinement), (c) holocomplex missing VHL (d) holocomplex missing CSN5/CSN6 and VHL, are shown in various orientations. (e) Resolution maps of (a-d) from front and back views. Map resolutions generated using RELION[231] software.



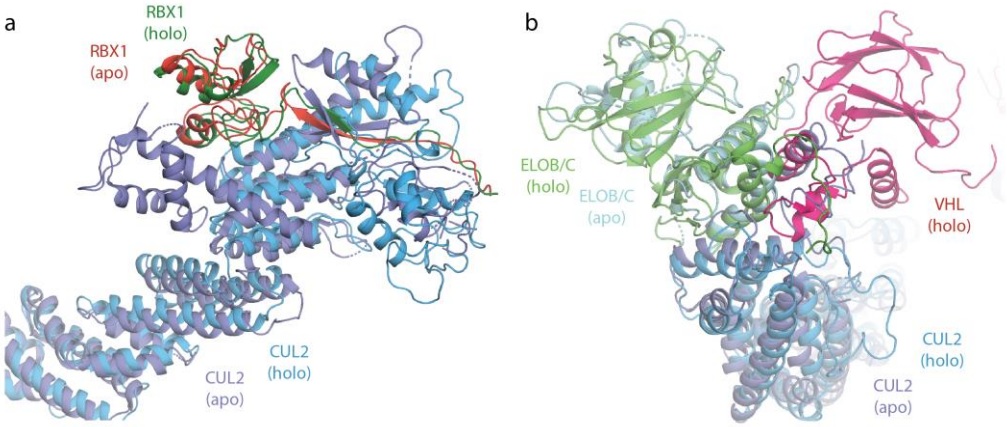
Supplementary Figure 6.22. Stereo images of CSN-CRL2~N8 complexes. Maps of the (a) CSN-CRL2~N8 holocomplex (threshold 0.022), (b) holocomplex (from VHL focused refinement; threshold 0.014), (c) holocomplex missing VHL (threshold 0.014) and (d) holocomplex missing CSN5/CSN6 and VHL (threshold 0.014).



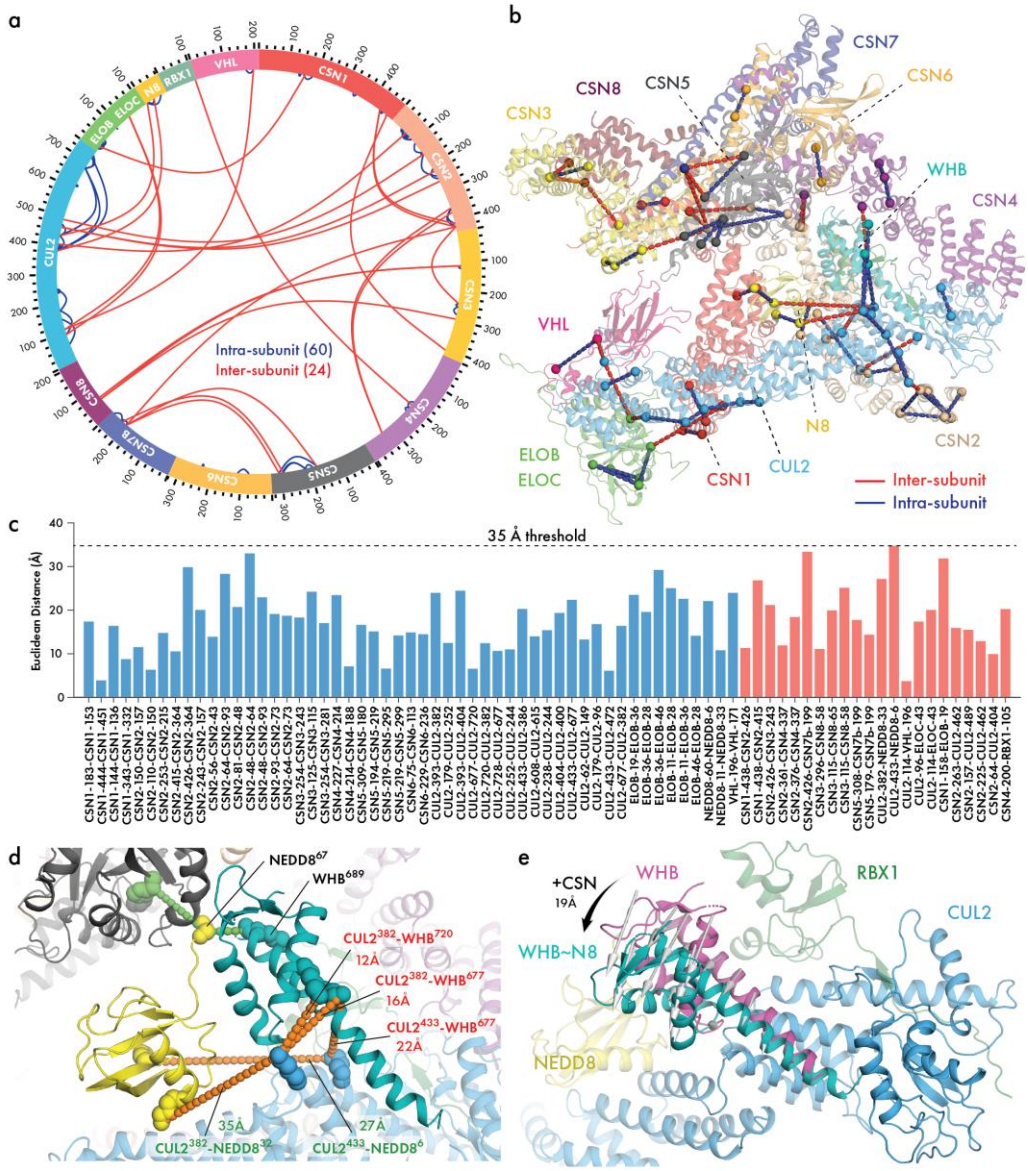
Supplementary Figure 6.23. Native MS of the CSN-CRL2~N8 complex. Spectra were collected from a 1:1 ratio of CSN and CRL2~N8, following a 1-hour incubation at room temperature and de-salting into 150 mM ammonium acetate (pH 7.5). The mass of each species were assigned using Waters MassLynx (4.1). The identity, charge, observed mass and percentage mass difference (compared to the expected mass) is shown for each complex. CSN3 and CSN5 subunits included a 2x StrepII and 6HIS N-terminal tags, respectively. The mass of all CSN complexes identified in the spectra include the tag masses of CSN3 and CSN5. The presence of the VBC complex on CRL2 has been explicitly marked, including any missing subunits from subcomplexes.



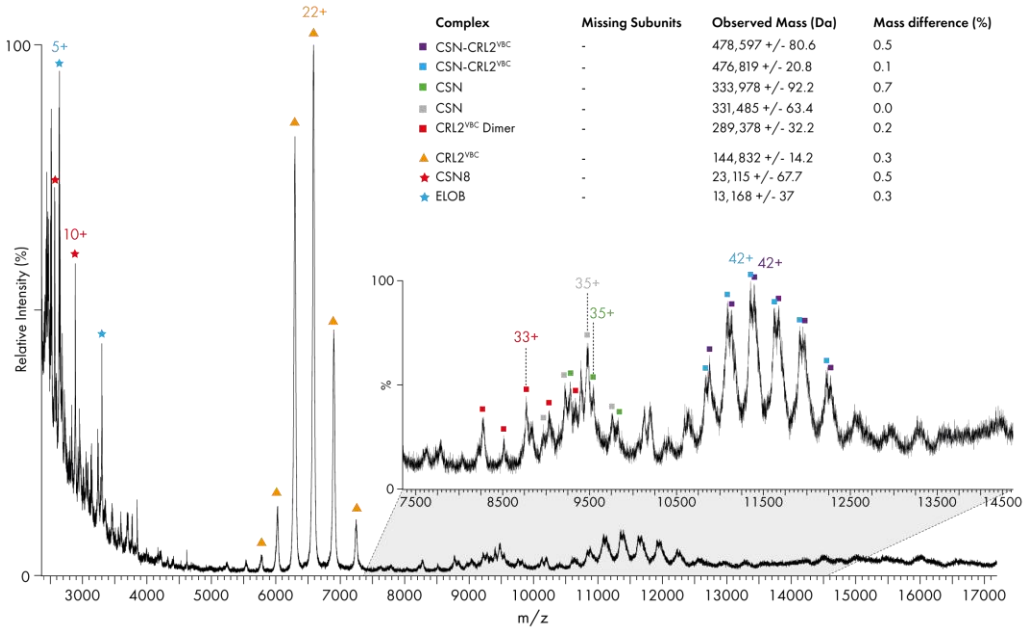
Supplementary Figure 6.24. Conformations of CSN2 and CSN4 in apo and holo CSN. (a-i) Conformations and crystal contacts of CSN2 and CSN4 for the nine independent molecules of apo-CSN in PDBs 4WSN, 4D10 and 4D18. Crystal contacts of CSN2 and CSN4 and neighbouring asymmetric units within 5 Å are shown by the green mesh. (j) Overlay of apo-CSN molecules in (a-i) showing range of CSN2 and CSN4 conformations. (k) Overlay of holo-CSN molecules in cryo-EM structures of CSN-CRL1 (dark blue), CSN-CRL2 (teal) and CSN-CRL4 (light blue). (l) Overlay of (j) and (k). CSN2 and CSN4 in apo and holo structures of the CSN represent two distinct clusters of conformations.



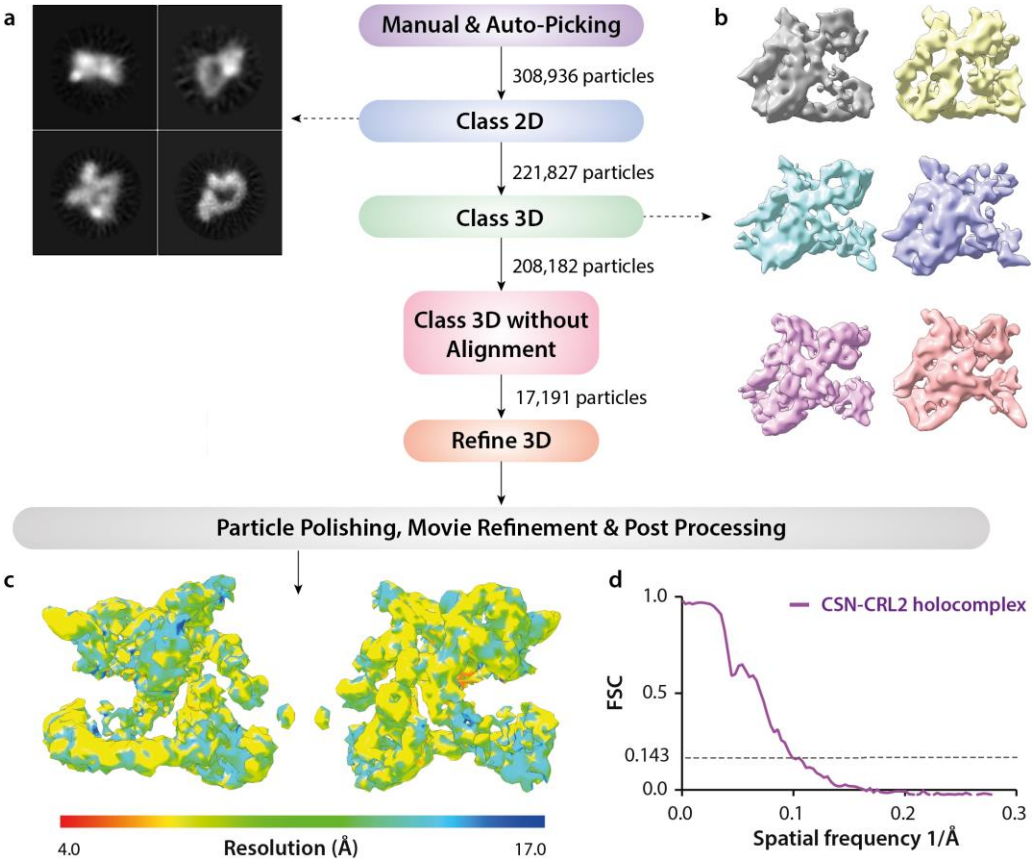
Supplementary Figure 6.25. Structural alignment of the CRL2. (a) C-terminal domain and (b) N-terminal domain in isolated (apo; 5N4W) and CSN-associated (holo) conformations. The structure of CSN has been hidden for clarity.



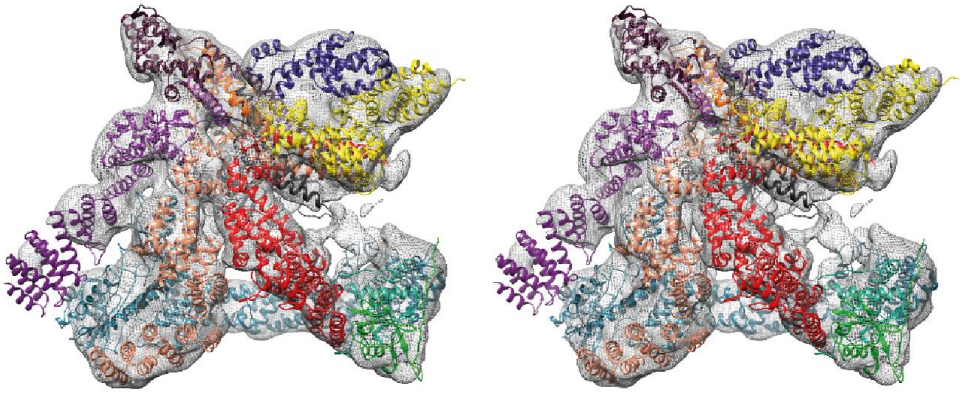
Supplementary Figure 6.26. Cross-links of the CSN-CRL2~N8 complex. (a) Circular plot of cross-links identified for the CSN-CRL2~N8 complex. (b) Inter- (red) and intra-subunit (blue) cross-links projected onto the cryo-EM structure of CSN-CRL2~N8. Nine cross-links were omitted due to missing residues in the C-terminal loops of CSN subunits (six) and self-residue cross-links (three) (**Supplementary Data 6.2**). (c) Euclidean distances for cross-links shown in (b). All measurements made between lysine Nz-Nz atoms using PyMOL. 35 Å cross-link distance threshold is shown by the dotted line. 35 Å takes into account two lysine sidechains (15 Å), 10 Å cross-linker length for BS3 and an additional 10 Å to account for domain-level flexibility. (d) Zoom of the WHB~N8 region of the CSN-CRL2~N8 model generated from cryo-EM and cross-link. The WHB domain is shown in teal, CUL2 in light blue (transparent), NEDD8 in yellow and CSN5 in grey. (e) Comparison of WHB~N8 from our hybrid cryo-EM/cross-linking CSN-CRL2~N8 model (teal) with non-neddylated WHB from CRL2 crystal structure (5N4W; purple). Vector arrows depict the 19 Å movement of WHB following neddylation and binding to CSN.



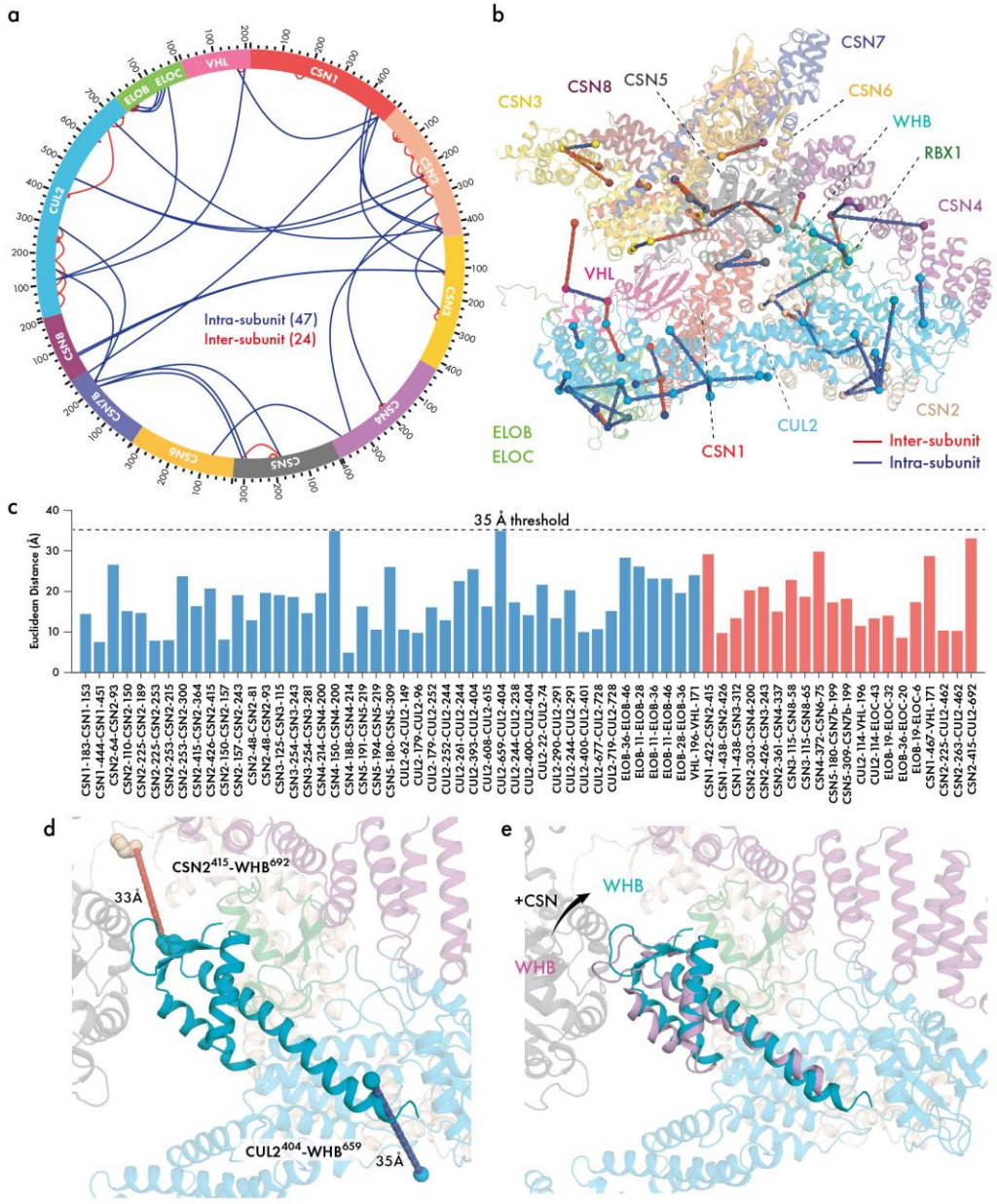
Supplementary Figure 6.27. Native MS of the CSN^{WT}-CRL2.



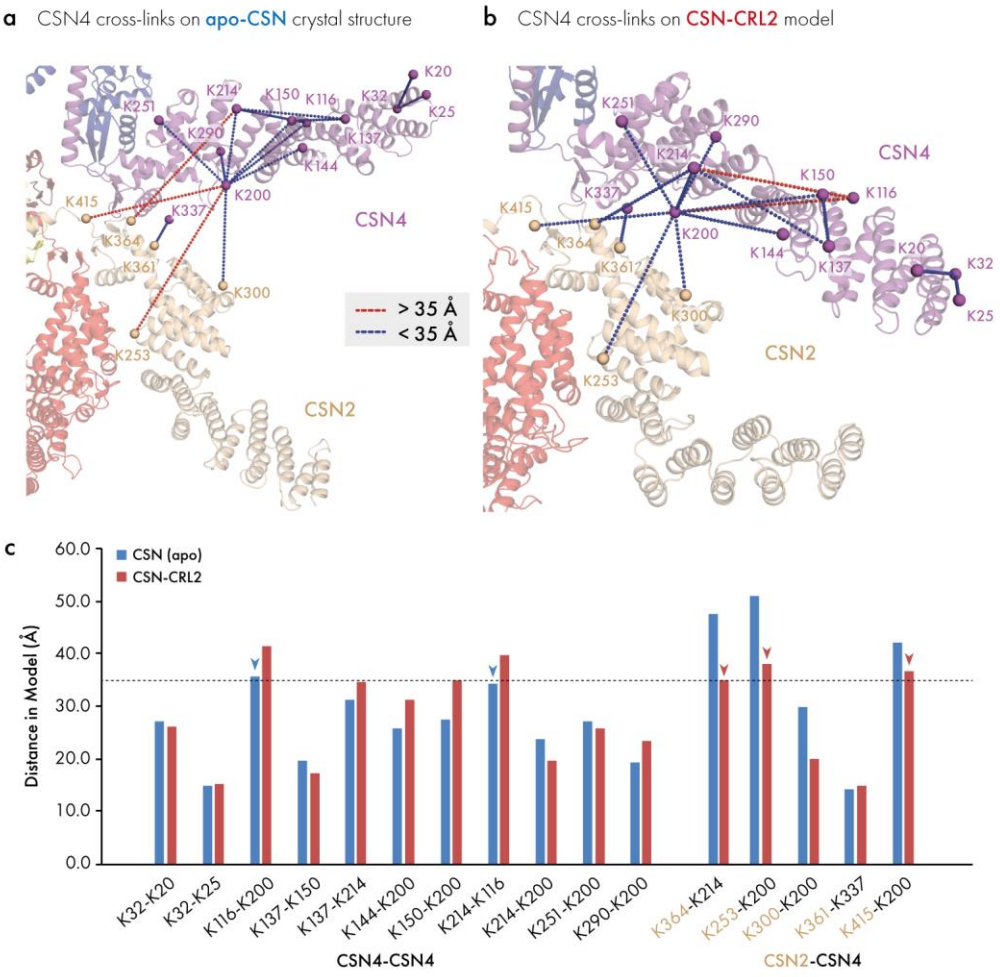
Supplementary Figure 6.28. Cryo-EM map of the CSN-CRL2 complex. A set of 6800 micrographs were subjected to manual and auto-picking in order to acquire particles for 2D reference-free classification (a). 2D classification was used for the positive selection of particles prior to 3D classification. (b) six classes generated, demonstrate subunit heterogeneity in the data set. (c) Resolution map of the CSN-CRL2 was generated using RELION[231]. (d) Fourier shell correlations (FSC) for the single CSN-CRL2 map.



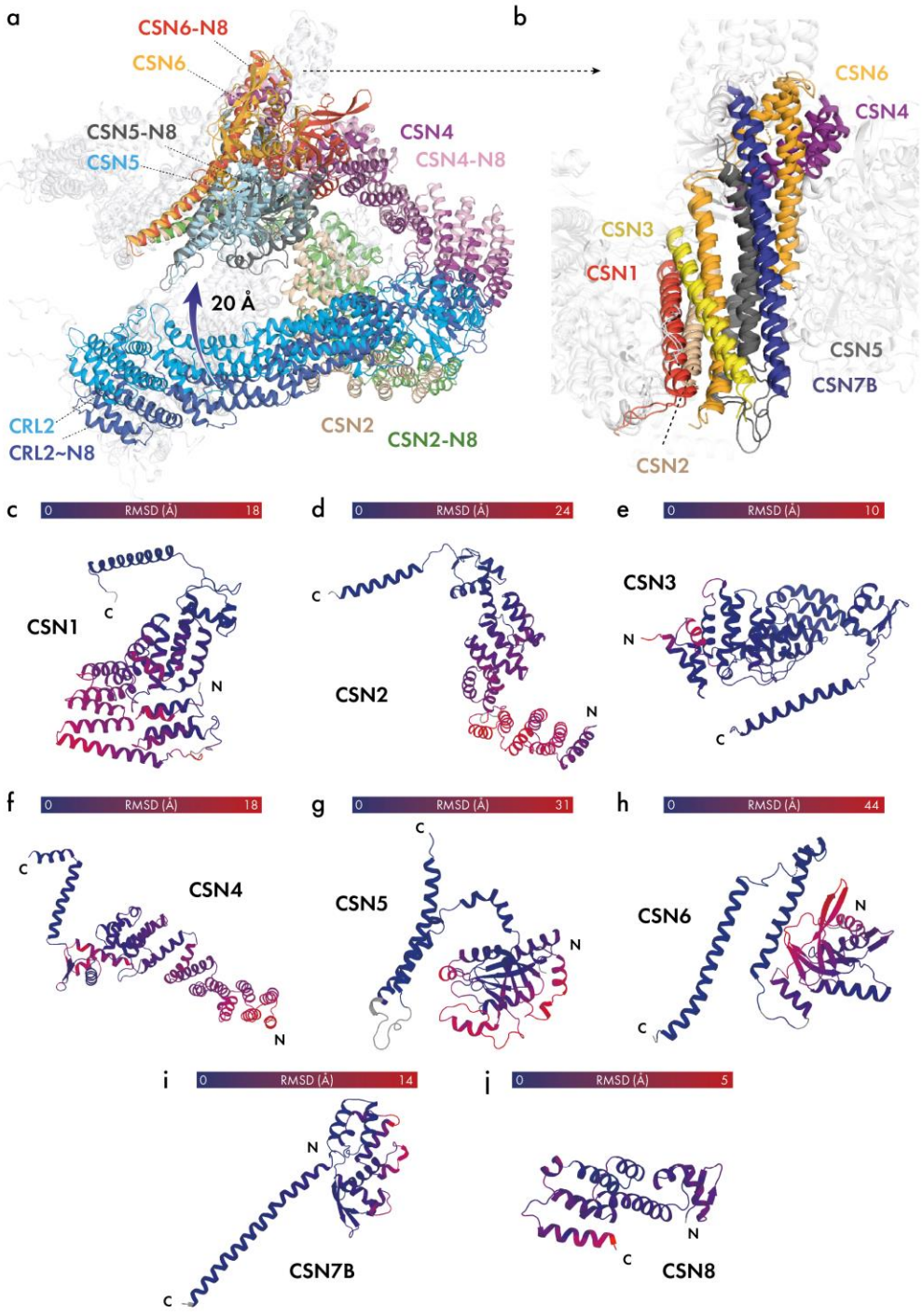
Supplementary Figure 6.29. Stereo images of the CSN-CRL2 complex. Threshold = 0.014.



Supplementary Figure 6.30. Cross-links of the deneddylated CSN-CRL2 complex. (a) Circular plot of cross-links identified for the CSN-CRL2 complex. (b) Inter- (red) and intra-subunit (blue) cross-links projected onto the cryo-EM structure of CSN-CRL2. Six cross-links were omitted due to missing residues in the C-terminal loops of CSN subunits (five) and self-residue cross-links (one) (**Supplementary Data 6.3**). (c) Euclidean distances for cross-links shown in (b). All measurements made between lysine Nz-Nz atoms using PyMOL. 35 Å cross-link distance threshold is shown by the dotted line. 35 Å takes into account two lysine sidechains (15 Å), 10 Å cross-linker length for BS3 and an additional 10 Å to account for domain-level flexibility. (e) Comparison of deneddylated WHB from our hybrid cryo-EM/cross-linking CSN-CRL2 model (teal) with WHB from CRL2 crystal structure (5N4W; purple).

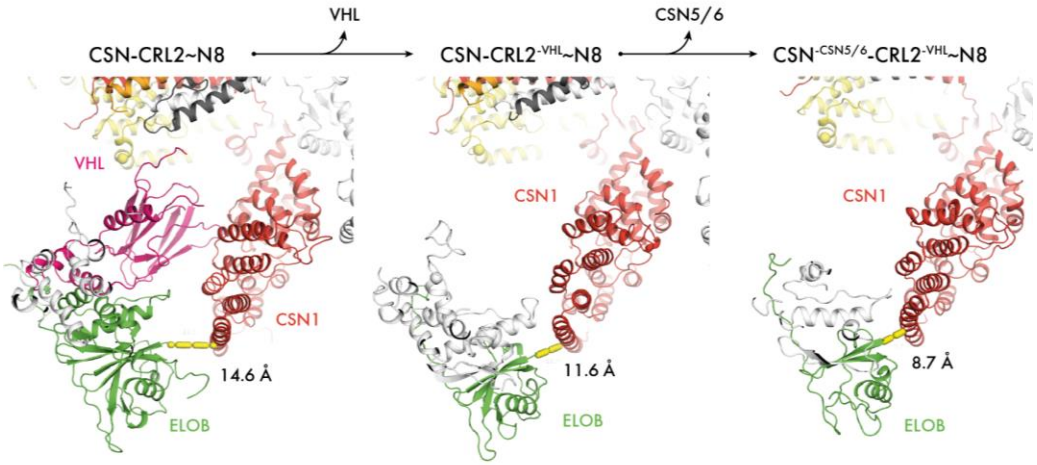


Supplementary Figure 6.31. Cross-links of CSN4 from apo-CSN^{WT}. Location and distances of CSN4 cross-links in apo-CSN^{WT}, projected onto the structures of (a) apo-CSN crystal structure (PDB 4D10) and (b) cryo-EM map fitted model of CSN-CRL2. Cross-links with distances satisfied under the cut-off of 35 Å are shown by dashed red lines, while those that are satisfied are in blue. (c) Bar plot showing distances of cross-links from (a-b). Dashed line represents a 35 Å distance threshold. Arrows mark cross-links which are satisfied in one conformation of the CSN but not the other. For K116-K200, K364-K214, K253-K200 and K415-K200 cross-links we included the shortest distance model in the satisfied category due to the distances being close to 35 Å, while the alternative conformation is much greater than 35 Å. All measurements were from lysine NZ atoms. The three CSN2-CSN4 cross links which differentiate between open and closed models (red arrows) are each located relatively close to the hinge domain of CSN4, nevertheless the distance between lysine residues changes quite substantially. Although potential cross links between residues closer to the N-termini of CSN2 and CSN4 would undergo a significantly greater change in separation, they were not observed and would not be expected to occur with the cross-linking reagent used here since the separation is substantially greater than 35 Å for both open and closed conformations.

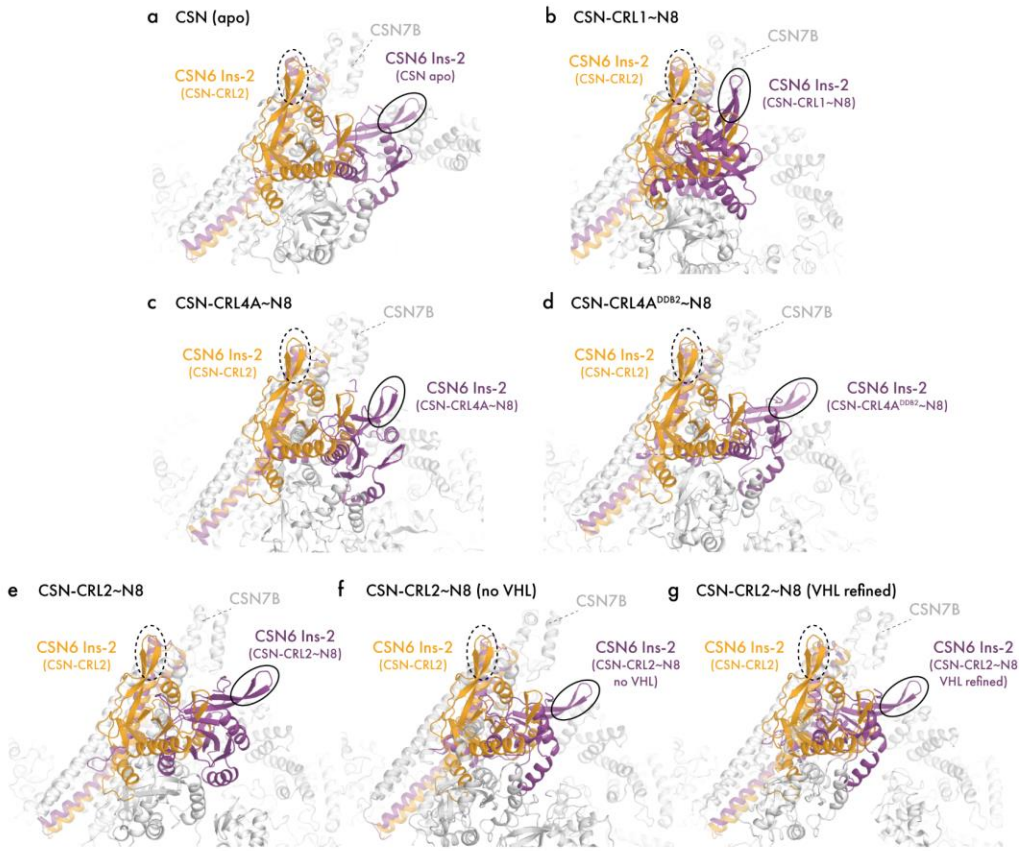


Supplementary Figure 6.32. Per-subunit comparisons of the CSN in neddylated and non-neddylated CSN-CRL2 complexes. (a) Superposition of the CSN-CRL2~N8 and CSN-CRL2. CSN1, CSN3, CSN7B, CSN8, RBX1, ELOB/C and VHL are shown in white to highlight the changes in CSN5/CSN6, CSN2/CSN4 and Cullin-2. Cullin-2 rotates upwards by ~20 Å in the absence of NEDD8. (b) Alignment of the C-terminal helical bundle of CSN-CRL2 and CSN-CRL2~N8 complexes. Per-subunit comparisons of (c) CSN1, (d) CSN2, (e) CSN3, (f), CSN4, (g) CSN5, (h) CSN6, (i) CSN7B, (j)

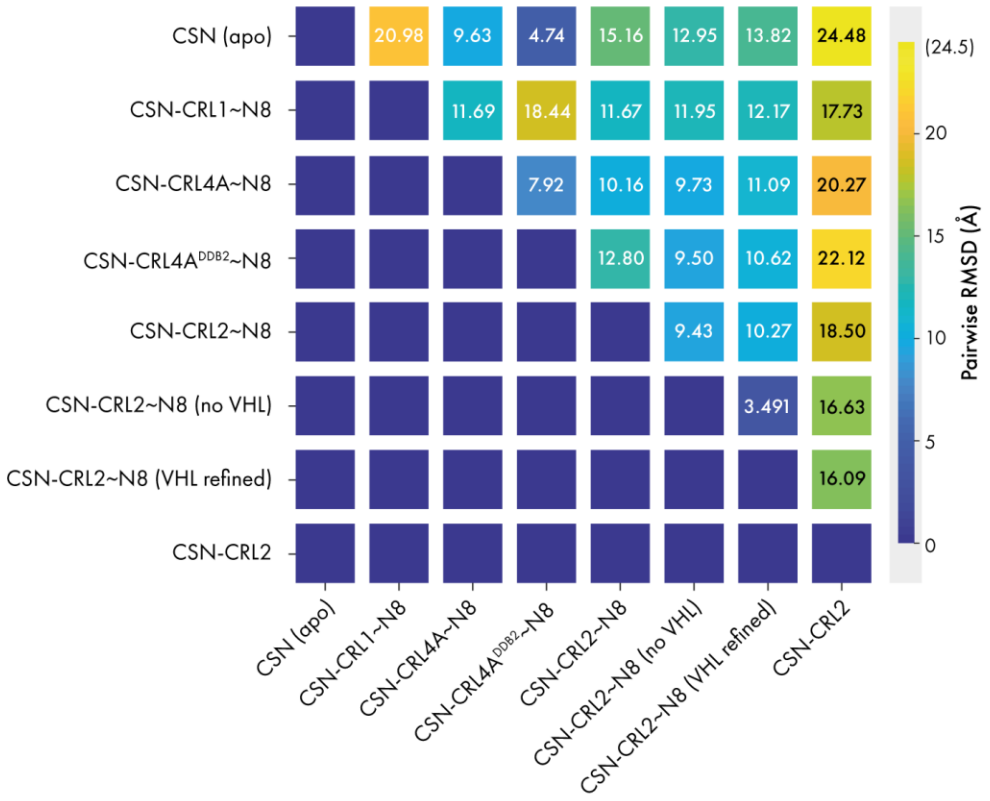
CSN8 between CSN-CRL2~N8 and CSN-CRL2 structures. The per-residue RMSD of each CSN1-8 subunit between CSN-CRL2~N8 and CSN-CRL2 complexes are indicated by the blue-red gradient. Each colour bar and colour gradient has been normalised to the maximum RMSD calculated for that subunit. The structure shown is the non-neddylated CSN-CRL2. The maximum per-subunit RMSDs measured were: CSN1 17.6 Å, CSN2 23.6 Å, CSN3 9.6 Å, CSN4 17.6 Å, CSN5 30.6 Å, CSN6 44.3 Å, CSN7B 13.5 Å, CSN8 5 Å.



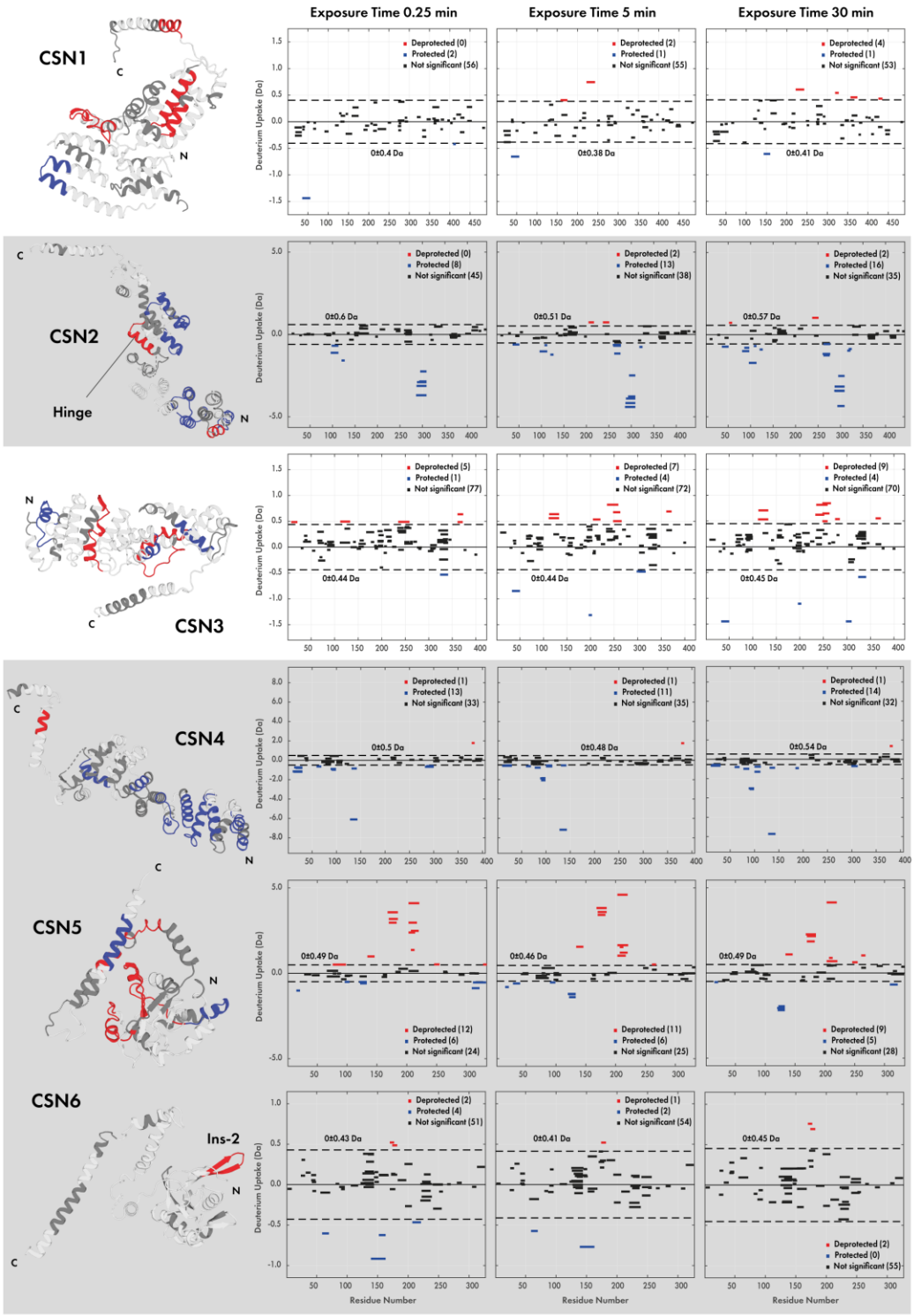
Supplementary Figure 6.33. CSN1-ELOB interface in CSN-CRL2~N8 complexes. The distance between CSN1 and ELOB of CSN-CRL2~N8, after loss of VHL and after loss of CSN5/CSN6 is shown by the dashed yellow line. All other CSN-CRL2 subunits have been coloured in white for clarity.

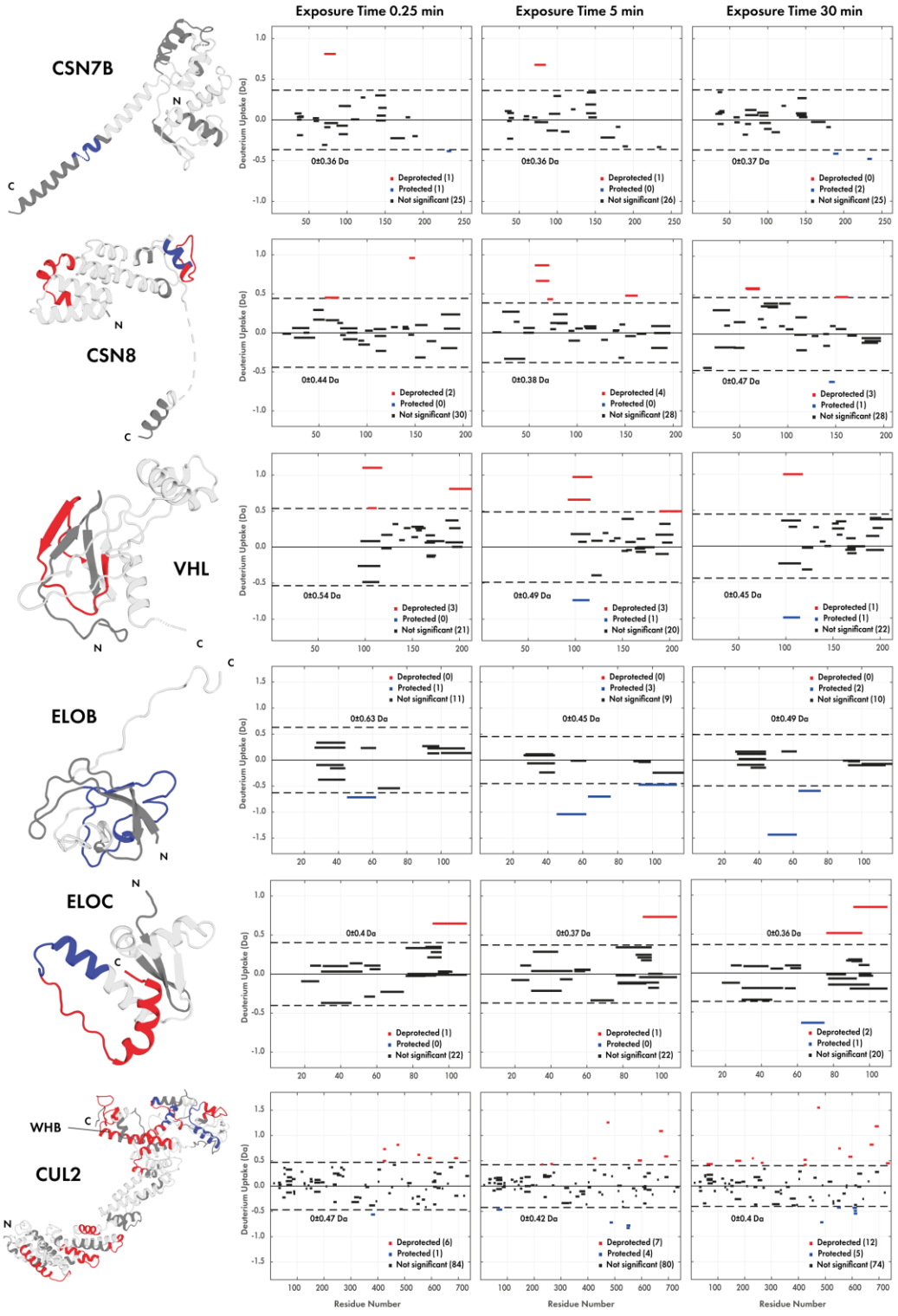


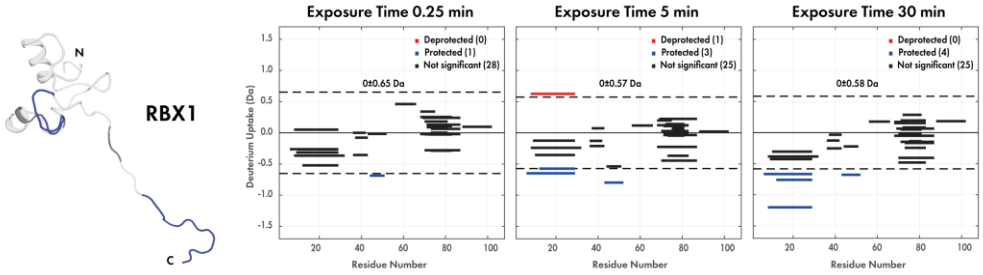
Supplementary Figure 6.34. Comparison of CSN6 conformations in published CSN and CSN-CRL complexes. Alignment of CSN-CRL2 (orange) with (a) crystal structure of apo CSN (PDB 4D10), (b) CSN-CRL1~N8 (EMD-3401), (c-d) fitted coordinates of CSN-CRL4A~N8 (EMD-3315) and CSN-CRL4A^{DDB2}~N8 (EMD-3316), (e-g) fitted coordinates of neddylated CSN-CRL2~N8 intact complex, missing VHL and VHL-refined (all purple). The CSN6 Ins-2 loop for each CSN6 has been highlighted for clarity (CSN6 of CSN-CRL2 with dashed ellipses, CSN6 of all other complexes in solid ellipses). Numerical quantification of similarity between CSN6 conformations is presented in Supplementary Figure 6.35.



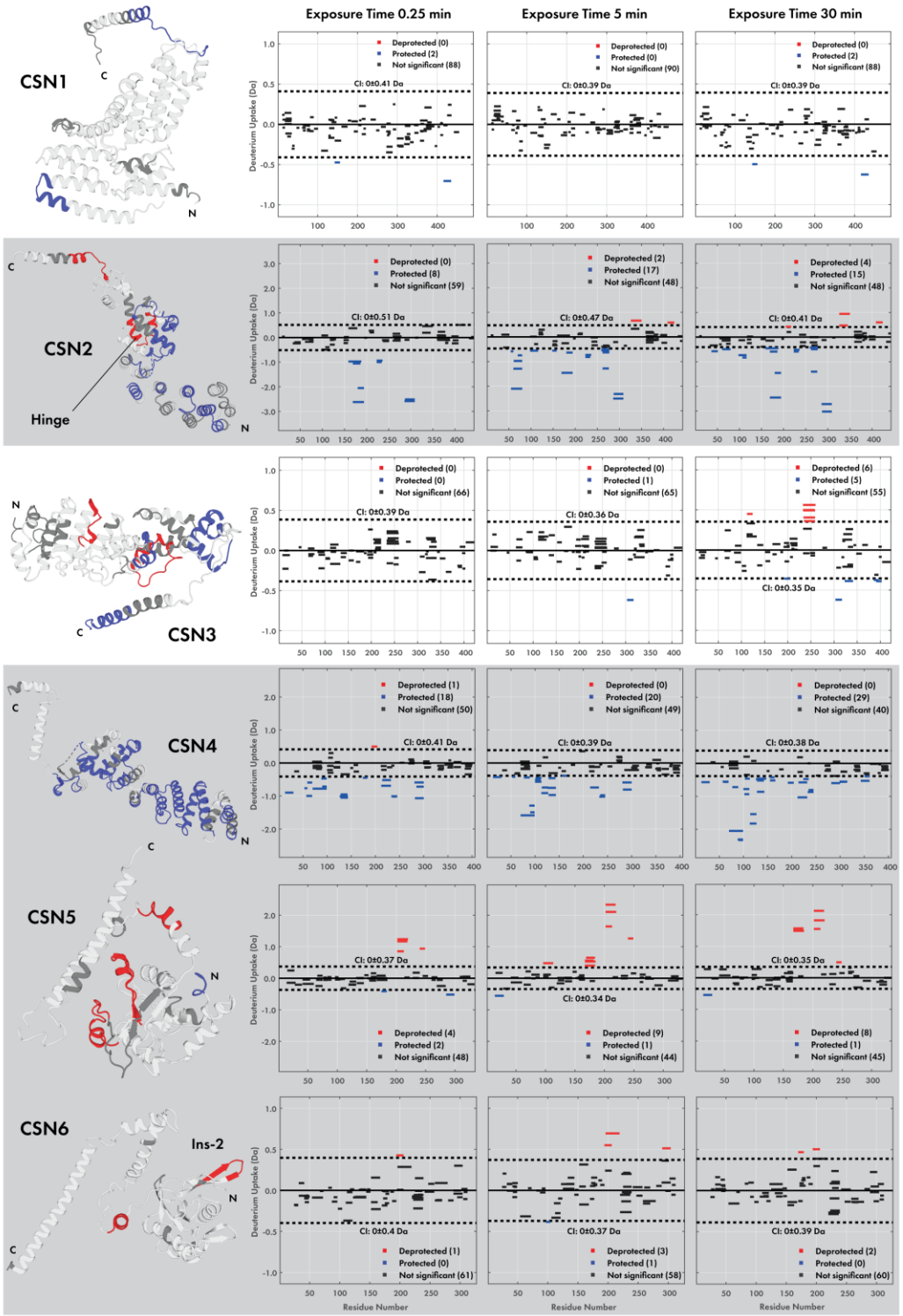
Supplementary Figure 6.35. Pairwise RMSD matrix of CSN6 in CSN and CSN-CRL complexes. Matrix indicates the RMSD between comparisons of CSN6 combinations in the apo CSN, CSN-CRL1~N8 (EMD-3401), CSN-CRL4A~N8 (EMD-3315), CSN-CRL4A^{DDB2}~N8 (EMD-3316), CSN-CRL2~N8, CSN-CRL2~N8 (-VHL), CSN-CRL2~N8 (VHL refined) and non-neddylated CSN-CRL2 complexes. The CSN1, CSN2, CSN3, CSN4, CSN7 and CSN8 subunits were aligned in each comparison and the RMSD was calculated between the non-fitted coordinates of CSN6 in each alignment using PyMOL. Aligned models are shown in Supplementary Figure 6.34.

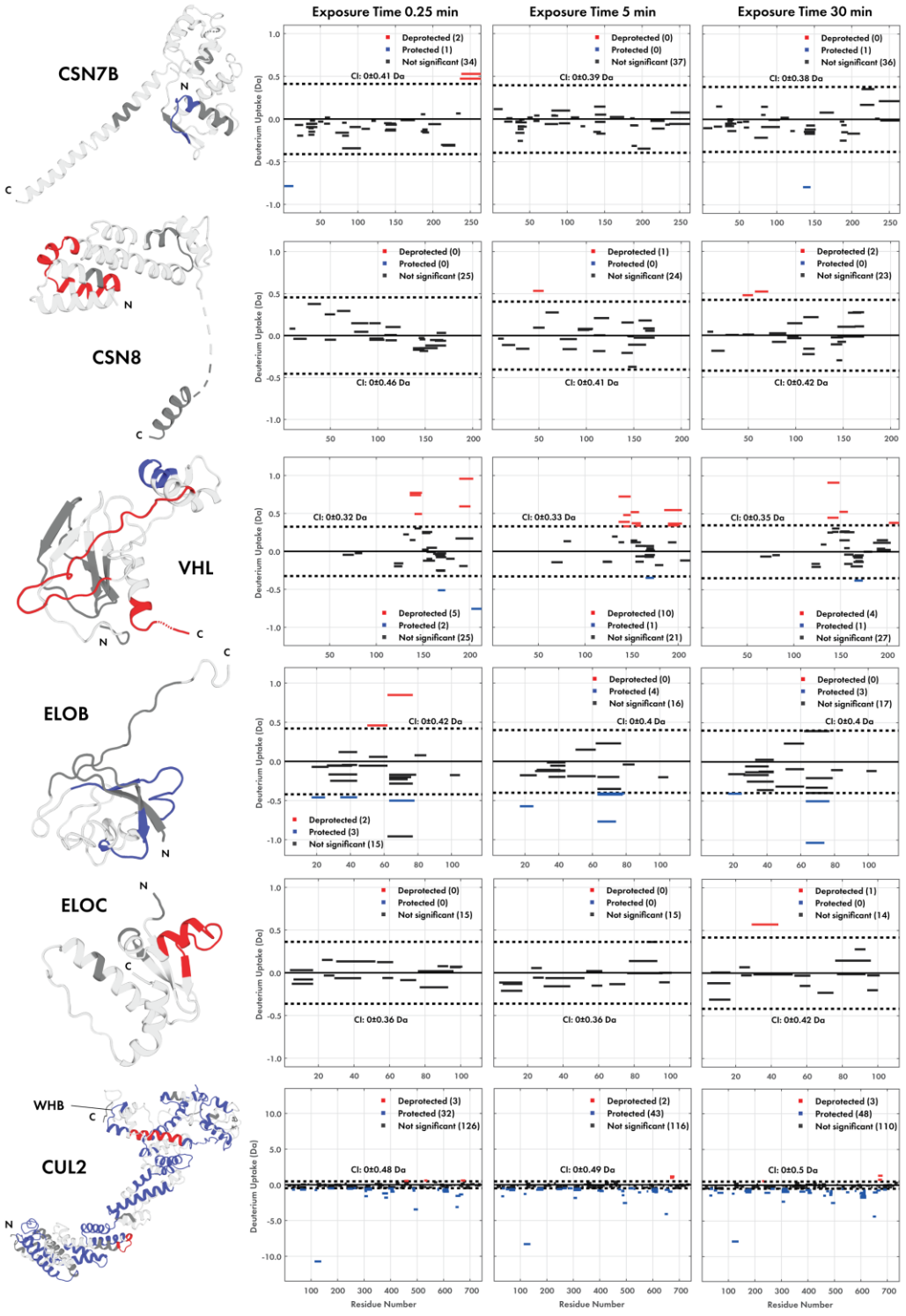

Exposure Time 0.25 min
Exposure Time 5 min
Exposure Time 30 min
CSN1
C
N
Deprotected (0)
Protected (2)
Not significant (56)
0±0.4 Da
Deprotected (2)
Protected (1)
Not significant (55)
0±0.38 Da
Deprotected (4)
Protected (1)
Not significant (53)
0±0.41 Da
CSN2
Hinge
Deprotected (0)
Protected (8)
Not significant (45)
0±0.6 Da
Deprotected (2)
Protected (13)
Not significant (38)
0±0.51 Da
Deprotected (2)
Protected (16)
Not significant (35)
0±0.57 Da
CSN3
Deprotected (5)
Protected (1)
Not significant (77)
0±0.44 Da
Deprotected (7)
Protected (4)
Not significant (72)
0±0.44 Da
Deprotected (9)
Protected (4)
Not significant (70)
0±0.45 Da
CSN4
Deprotected (1)
Protected (13)
Not significant (33)
0±0.5 Da
Deprotected (1)
Protected (11)
Not significant (35)
0±0.48 Da
Deprotected (1)
Protected (14)
Not significant (32)
0±0.54 Da
CSN5
Deprotected (12)
Protected (6)
Not significant (24)
0±0.49 Da
Deprotected (11)
Protected (6)
Not significant (25)
0±0.46 Da
Deprotected (9)
Protected (5)
Not significant (28)
0±0.49 Da
CSN6
Ins-2
Deprotected (2)
Protected (4)
Not significant (51)
0±0.43 Da
Deprotected (1)
Protected (2)
Not significant (54)
0±0.41 Da
Deprotected (2)
Protected (0)
Not significant (55)
0±0.45 Da
Deuterium Uptake (Da)
Residue Number

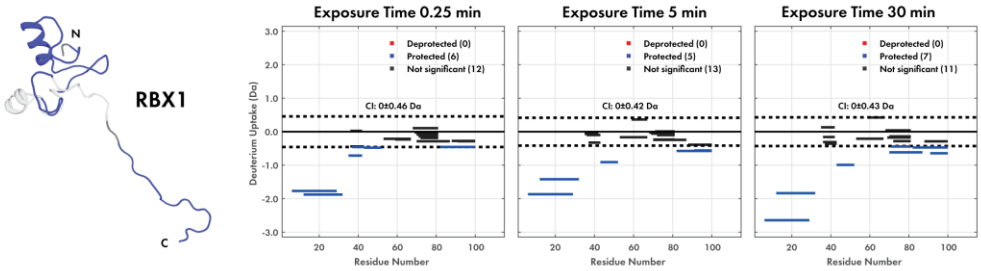

Exposure Time 0.25 min
Exposure Time 5 min
Exposure Time 30 min
CSN7B
CSN8
VHL
ELOB
ELOC
WHB
CUL2
N
C
Deuterium Uptake (Da)
Residue Number
0±0.36 Da
Deprotected (1)
Protected (1)
Not significant (25)
0±0.36 Da
Deprotected (1)
Protected (0)
Not significant (26)
0±0.37 Da
Deprotected (0)
Protected (2)
Not significant (25)
0±0.44 Da
Deprotected (2)
Protected (0)
Not significant (30)
0±0.38 Da
Deprotected (4)
Protected (0)
Not significant (28)
0±0.47 Da
Deprotected (3)
Protected (1)
Not significant (28)
0±0.54 Da
Deprotected (3)
Protected (0)
Not significant (21)
0±0.49 Da
Deprotected (3)
Protected (1)
Not significant (20)
0±0.45 Da
Deprotected (1)
Protected (1)
Not significant (22)
0±0.63 Da
Deprotected (0)
Protected (1)
Not significant (11)
0±0.45 Da
Deprotected (0)
Protected (3)
Not significant (9)
0±0.49 Da
Deprotected (0)
Protected (2)
Not significant (10)
0±0.4 Da
Deprotected (1)
Protected (0)
Not significant (22)
0±0.37 Da
Deprotected (1)
Protected (0)
Not significant (22)
0±0.36 Da
Deprotected (2)
Protected (1)
Not significant (20)
0±0.47 Da
Deprotected (6)
Protected (1)
Not significant (84)
0±0.42 Da
Deprotected (7)
Protected (4)
Not significant (80)
0±0.4 Da
Deprotected (12)
Protected (5)
Not significant (74)



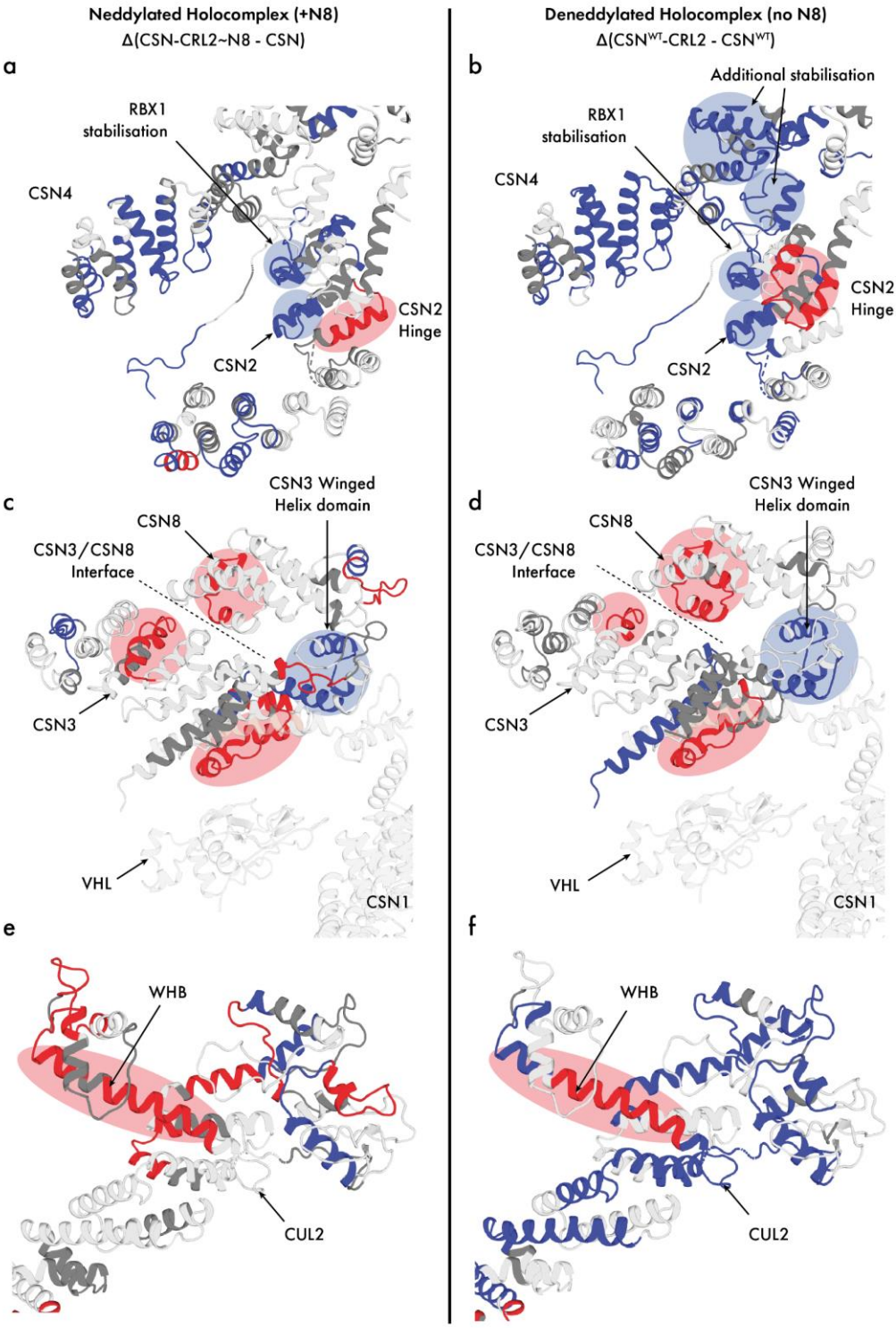
Supplementary Figure 6.36. HDX-MS of CSN-CRL2~N8 per protein per timepoint. Differential comparison of Δ(CSN-CRL2~N8 – CSN). Peptides experiencing stabilisation upon CRL2~N8 binding to CSN, compared to the peptide in apo-CSN, are shown as blue, destabilised peptides are shown in red. CSN2, CSN4, CSN5 and CSN6 have been highlighted by grey boxes. The CSN2 hinge and Cullin-2 WHB domain have been highlighted in grey for clarity. Peptides were filtered applying a 98% confidence limit (critical value of 6.965; dotted lines). Structures show data for the 30 min timepoint.


Exposure Time 0.25 min
Exposure Time 5 min
Exposure Time 30 min
CSN1
C
N
Deprotected (0)
Protected (2)
Not significant (88)
CI: 0±0.41 Da
Deprotected (0)
Protected (0)
Not significant (90)
CI: 0±0.39 Da
Deprotected (0)
Protected (2)
Not significant (88)
CI: 0±0.39 Da
CSN2
Hinge
Deprotected (0)
Protected (8)
Not significant (59)
CI: 0±0.51 Da
Deprotected (2)
Protected (17)
Not significant (48)
CI: 0±0.47 Da
Deprotected (4)
Protected (15)
Not significant (48)
CI: 0±0.41 Da
CSN3
Deprotected (0)
Protected (0)
Not significant (66)
CI: 0±0.39 Da
Deprotected (0)
Protected (1)
Not significant (65)
CI: 0±0.36 Da
Deprotected (6)
Protected (5)
Not significant (55)
CI: 0±0.35 Da
CSN4
Deprotected (1)
Protected (18)
Not significant (50)
CI: 0±0.41 Da
Deprotected (0)
Protected (20)
Not significant (49)
CI: 0±0.39 Da
Deprotected (0)
Protected (29)
Not significant (40)
CI: 0±0.38 Da
CSN5
Deprotected (4)
Protected (2)
Not significant (48)
CI: 0±0.37 Da
Deprotected (9)
Protected (1)
Not significant (44)
CI: 0±0.34 Da
Deprotected (8)
Protected (1)
Not significant (45)
CI: 0±0.35 Da
CSN6
Ins-2
Deprotected (1)
Protected (0)
Not significant (61)
CI: 0±0.4 Da
Deprotected (3)
Protected (1)
Not significant (58)
CI: 0±0.37 Da
Deprotected (2)
Protected (0)
Not significant (60)
CI: 0±0.39 Da
Deuterium Uptake (Da)
Residue Number


Exposure Time 0.25 min
Exposure Time 5 min
Exposure Time 30 min
CSN7B
CSN8
VHL
ELOB
ELOC
WHB
CUL2
Deuterium Uptake (Da)
Residue Number
Deprotected (2)
Protected (1)
Not significant (34)
CI: 0±0.41 Da
Deprotected (0)
Protected (1)
Not significant (37)
CI: 0±0.39 Da
Deprotected (0)
Protected (1)
Not significant (36)
CI: 0±0.38 Da
Deprotected (0)
Protected (0)
Not significant (25)
CI: 0±0.46 Da
Deprotected (1)
Protected (0)
Not significant (24)
CI: 0±0.41 Da
Deprotected (2)
Protected (0)
Not significant (23)
CI: 0±0.42 Da
Deprotected (5)
Protected (2)
Not significant (25)
CI: 0±0.32 Da
Deprotected (10)
Protected (1)
Not significant (21)
CI: 0±0.33 Da
Deprotected (4)
Protected (1)
Not significant (27)
CI: 0±0.35 Da
Deprotected (2)
Protected (3)
Not significant (15)
CI: 0±0.42 Da
Deprotected (0)
Protected (4)
Not significant (16)
CI: 0±0.4 Da
Deprotected (0)
Protected (3)
Not significant (17)
CI: 0±0.4 Da
Deprotected (0)
Protected (0)
Not significant (15)
CI: 0±0.36 Da
Deprotected (0)
Protected (0)
Not significant (15)
CI: 0±0.36 Da
Deprotected (1)
Protected (0)
Not significant (14)
CI: 0±0.42 Da
Deprotected (3)
Protected (32)
Not significant (126)
CI: 0±0.48 Da
Deprotected (2)
Protected (43)
Not significant (116)
CI: 0±0.49 Da
Deprotected (3)
Protected (48)
Not significant (110)
CI: 0±0.5 Da

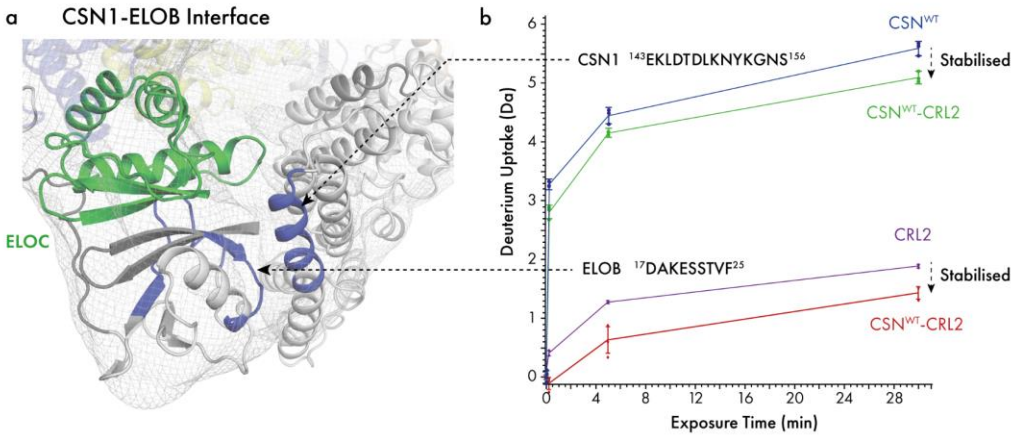


Supplementary Figure 6.37. HDX-MS of CSN^{WT}-CRL2 per protein per timepoint. Differential comparison of Δ(CSN^{WT}-CRL2 – CSN^{WT}). Peptides experiencing stabilisation upon non-neddylated CRL2 binding to CSN^{WT}, compared to the peptide in apo-CSN^{WT}, are shown as blue, destabilised peptides are shown in red. CSN2, CSN4, CSN5 and CSN6 have been highlighted by grey boxes. The CSN2 hinge and Cullin-2 WHB domain have been highlighted for clarity. Peptides were filtered applying a 98% confidence limit (critical value of 6.965; dotted lines). Structures show data for the 30 min timepoint.

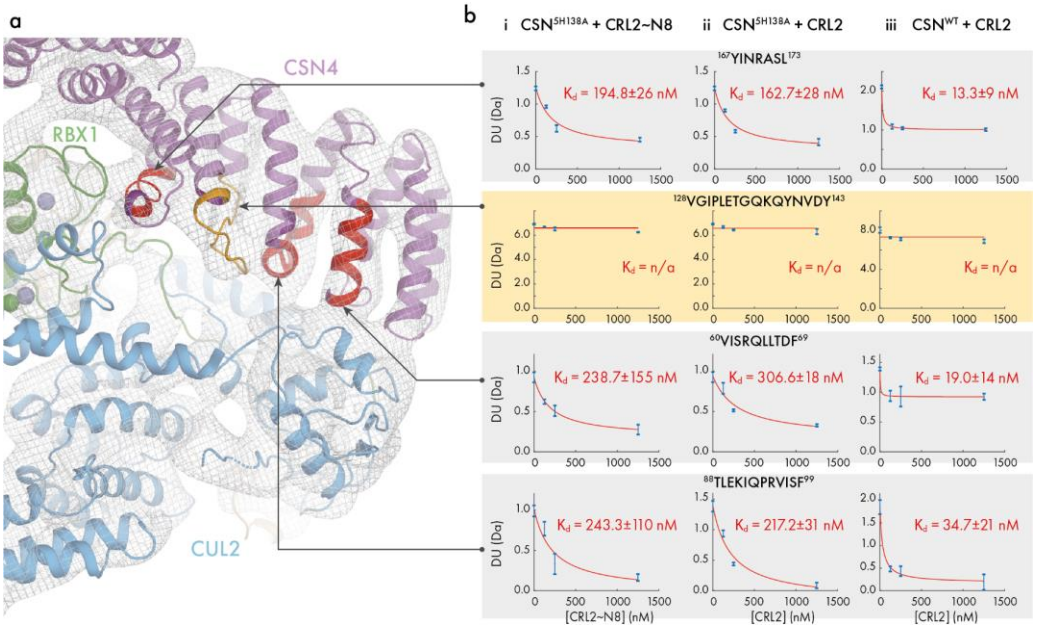


Supplementary Figure 6.38. ΔHDX changes in CSN2/CSN4/RBX1, CSN3/CSN8 and CUL2. Regions showing significant destabilisation (red) and stabilisation (blue) are shown on the structures of CSN-CRL2~N8 (left) and CSN-CRL2 (right) and highlighted for clarity. (a-b) Stabilisation of CSN2 and RBX1 interfaces. In the deneddylated complex, CSN4 and RBX1 stabilisation suggests an

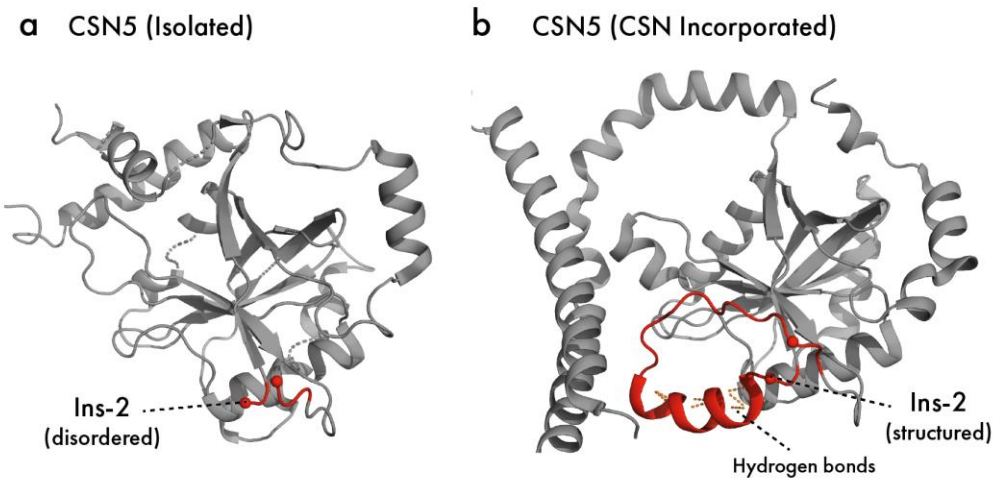
interface between the two subunits (b; top right blue circles). (c-d) HDX changes in CSN3, CSN8. VHL and CSN1 (shown in white) are displayed for reference. Both CSN3 and CSN8 experience destabilisation at their interfaces in both left and right complexes. The CSN3 surface closest to VHL (red) and the CSN3 winged helix domain (blue) also shows consistent differences. (e-f) Destabilisation of the WHB domain Cullin-2 (CUL2) in both conditions.



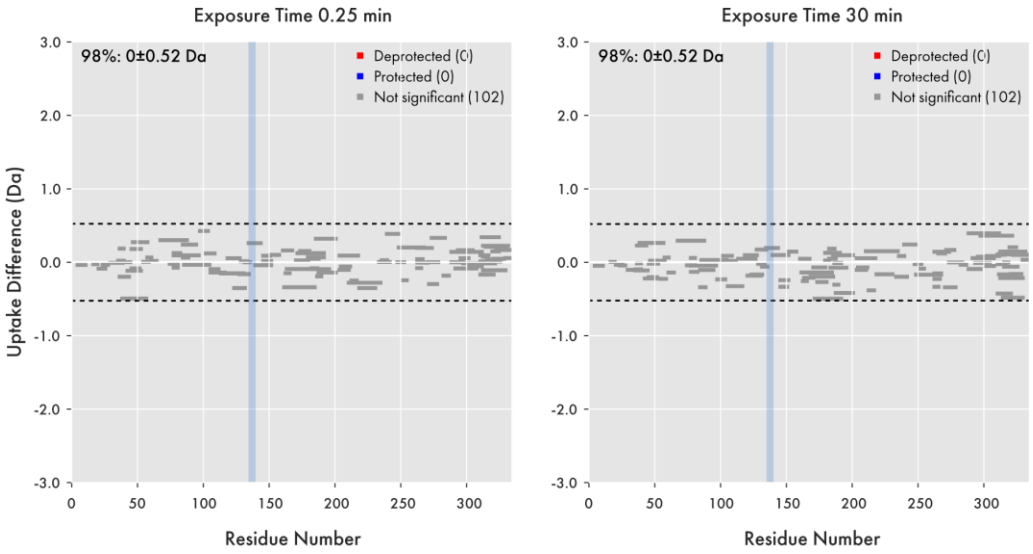
Supplementary Figure 6.39. CSN1-ELOB interface of CSN^{WT}-CRL2 in HDX-MS. (a) Structure of CSN-CRL2 fitted to cryo-EM density. Significantly stabilised peptides from CSN1 and ELOB between its interface is highlighted in blue. (b) Deuterium uptake curves over 30 minutes for CSN1 and ELOB peptides are shown for the Δ(CSN^{WT}-CRL2 – CSN^{WT}) comparison. Error bars represent the deuterium uptake standard deviation.



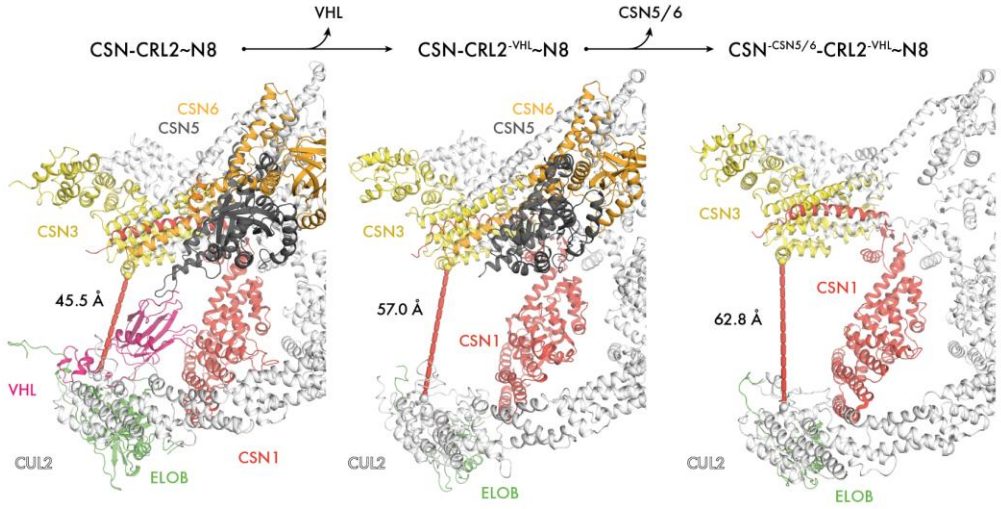
Supplementary Figure 6.40. Dissociation constants between CSN4 and CRL2/CRL2~N8 in CSN-CRL2 complexes. (a) Structure and density map of the intact CSN-CRL2~N8 shown for reference. Three peptides identified as interacting with CRL2/CRL2~N8 in PLIMSTEX experiments have been highlighted in red. An example of a peptide with no observed changes is shown in orange. (b) PLIMSTEX curve plots showing the deuterium uptake of CSN peptides from (i) CSN-CRL2~N8, (ii) CSN-CRL2 and (iii) CSN^{WT}-CRL2, as a function of increasing concentrations of either CRL2 or CRL2~N8. Data points represent the average deuterium uptake and error bars indicate standard deviation of technical triplicates. The red curve for interacting peptides was fitted using a 3-parameter 1:1 binding model for 250 nM CSN or CSN^{WT}, titrated with CRL2 or CRL2~N8 from 1:0 to 1:5 molar ratios. Kd ± values denote the standard deviation of the Kd measurement from technical triplicates.



Supplementary Figure 6.41. CSN5 Ins-2 loop in isolated and CSN incorporated structures. (a) Disordered Ins-2 loop in isolated CSN (PDB 4F7O). (b) Structured Ins-2 loop in CSN incorporated CSN5 (PDB 4D10). Hydrogen bonds of the Ins-2 helix have been shown in orange.



Supplementary Figure 6.42. Woods plot comparing deuterium uptake difference of CSN5 peptides from apo-CSN^{WT} and CSN^{5H138A} complexes. Each CSN5 peptide is represented by a horizontal bar and are coloured according to the significance of its uptake difference. Dotted line represents a 98% confidence interval used to filter peptides for statistical significance. All 102 peptides identified from CSN5 showed non-statistically significant changes in deuterium uptake. The position of the CSN5 H138A mutation has been highlighted by the blue box. Plots generated using Deuteros (v1.08).



Supplementary Figure 6.43. Conformational heterogeneity of the CSN-CRL2~N8 structures. The distance between CSN3-CUL2 N-terminal domain of CSN-CRL2~N8, after loss of VHL and after loss of CSN5/CSN6 is shown by the dashed red line. All other CSN-CRL2 subunits have been coloured in white for clarity.

6.3.2 Supplementary Tables

Supplementary Table 6.2. Kd values determined for CSN-CRL1 and CSN-CRL2 complexes

	CRL1[n]	CRL1~N8[n]	CRL2[o]	CRL2~N8[o]
CSN^{WT}	310.0	-	22.3	-
CSN^{5H138A}	10.0	1.6	228.8	225.6

n Values accessed from Mosadeghi *et al*. 2016 ref 222

o Average Kd values from PLIMSTEX

6.3.3 Supplementary Notes

Supplementary Note 6.1. Multi-template homology modelling of CRL2 using MODELLER.

```
from modeller.automodel import *

log.verbose()    # request verbose output
env = environ(rand_seed=-556)

# Read in HETATM records from template PDBs
env.io.hetatm = True

a = automodel(env, alnfile='alignment.ali',
              knowns=('5N4W_noWHA','4WQO','CRL2_1LDJ_model'),
              sequence='CRL2',
              assess_methods=(assess.DOPE, assess.GA341))

a.initial_malign3d = False
a.starting_model = 1
a.ending_model = 1
a.md_level = refine.slow
a.make()

# Run script using 'mod9.16 model_mult.py'

```

Supplementary Note 6.2. IMP XL-modelling script for position of subunits.

```
import IMP
import IMP.core
import IMP.algebra
import IMP.atom
import IMP.container

import IMP.pmi.restraints.crosslinking
import IMP.pmi.restraints.stereochemistry
import IMP.pmi.restraints.em
import IMP.pmi.restraints.basic
import IMP.pmi.representation
import IMP.pmi.tools
import IMP.pmi.samplers
import IMP.pmi.output
import IMP.pmi.macros
import IMP.pmi.topology

import os
import sys
import csv

# Define Input Files
datadirectory = "./inputs/"
topology_file = datadirectory+"topology.txt"

# Set MC Sampling Parameters
num_frames = 1000
if '--test' in sys.argv: num_frames=50
num_mc_steps = 10
rb_max_trans = 2.00
rb_max_rot = 0.1
bead_max_trans = 0.05

rigid_bodies = [["CUL2-WHB"],["VHL"],['NEDD8']]

# Build the Model Representation
m = IMP.Model()

# Create list of components from topology file
topology = IMP.pmi.topology.TopologyReader(topology_file)
domains = topology.component_list

print('#'*10,domains)

bm = IMP.pmi.macros.BuildModel(m,
                    component_topologies=domains,
```

```
                    list_of_rigid_bodies=rigid_bodies)

representation = bm.get_representation()

# add colors to the components
for nc,component in enumerate(domains):
    name = component.name
    sel = IMP.atom.Selection(representation.prot,molecule=name)
    ps = sel.get_selected_particles()
    clr = IMP.display.get_rgb_color(float(nc)/len(domains))
    for p in ps:
        if not IMP.display.Colored.get_is_setup(p):
            IMP.display.Colored.setup_particle(p,clr)
        else:
            IMP.display.Colored(p).set_color(clr)

# Define Degrees of Freedom
representation.set_rigid_bodies_max_rot(rb_max_rot)
representation.set_floppy_bodies_max_trans(bead_max_trans)
representation.set_rigid_bodies_max_trans(rb_max_trans)

outputobjects = []
sampleobjects = []

outputobjects.append(representation)
sampleobjects.append(representation)

# Excluded Volume Restraint
ev = IMP.pmi.restraints.stereochemistry.ExcludedVolumeSphere(
                                        representation, resolution=10)
ev.add_to_model()
outputobjects.append(ev)


# Crosslinks - dataset 1
columnmap={}
columnmap["Protein1"]="prot1"
columnmap["Protein2"]="prot2"
columnmap["Residue1"]="res1"
columnmap["Residue2"]="res2"
columnmap["IDScore"]=None

# Experimentally measured crosslinks
xl1 = IMP.pmi.restraints.crosslinking.ISDCrossLinkMS(representation,
                                    datadirectory+'xlinks_exp.txt',
                                    length=35.0,
                                    slope=0.1,
                                    columnmapping=columnmap,
                                    resolution=1.0,
                                    label="Inter-crosslinks",
                                    csvfile=True)
xl1.add_to_model()
sampleobjects.append(xl1)
outputobjects.append(xl1)
```

```
101
102 # Pseudocovalent crosslinks for keeping complex integrity and allowing
103 # flexibility of long flexible domains
104 xl2 = IMP.pmi.restraints.crosslinking.ISDCrossLinkMS(representation,
105                     datadirectory+'xlinks_rigidbody_floppy_pseudocovalent.txt',
106                     length=5.0,
107                     slope=1,
108                     columnmapping=columnmap,
109                     resolution=1.0,
110                     label="primary_covalents",
111                     csvfile=True)
112
113 xl2.add_to_model()
114 sampleobjects.append(xl2)
115 outputobjects.append(xl2)
116
117 # Pseudocovalent crosslinks for keeping complex integrity and allowing
118 # flexibility of long flexible domains
119 xl3 = IMP.pmi.restraints.crosslinking.ISDCrossLinkMS(representation,
120                     datadirectory+'xlinks_rigidbody_floppy_isopeptide.txt',
121                     length=3.0,
122                     slope=1,
123                     columnmapping=columnmap,
124                     resolution=1.0,
125                     label="primary_covalents",
126                     csvfile=True)
127
128 xl3.add_to_model()
129 sampleobjects.append(xl3)
130 outputobjects.append(xl3)
131
132 mc1=IMP.pmi.macros.ReplicaExchange0(m,
133                                     representation,
134                                     monte_carlo_sample_objects=sampleobjects,
135                                     output_objects=outputobjects,
136                                     monte_carlo_temperature=1.0,
137                                     crosslink_restraints=[xl1,xl2,xl3],
138                                     simulated_annealing=False,
139                                     number_of_best_scoring_models=100,
140                                     monte_carlo_steps=num_mc_steps,
141                                     number_of_frames=num_frames,
142                                     global_output_directory="output",
143                                     atomistic=True)
144
145 mc1.execute_macro() # start
146
147 # Run script using 'python CSN_XL_modeling.py'
148
```

6.3.4 Supplementary Data & Movie

Supplementary Data 6.1-Supplementary Data 6.3 each contain cross-link lists of the CSN-CRL2~N8, CSN-CRL2 and CSN complexes including the exact peptide sequence, number of spectra, max pLink score, calculated precursor mass and observed charge states. Supplementary Data files 6.1-6.3 are too large to be included in this thesis, hence a summarised list for each complex has been included here.

Supplementary Movie 6.1. The showcases conformational changes between the different CSN-CRL2 structures shown in this article and can be accessed Nature Communications website[p].

p Accessible from https://www.nature.com/articles/s41467-019-11772-y

Supplementary Data 6.1. Chemical cross-links of the CSN-CRL2~N8 complex.

Index	Protein A	Protein B	Residue A	Residue B
Intra-subunit				
1	CSN1	CSN1	183	153
2	CSN1	CSN1	444	451
3	CSN1	CSN1	144	136
4	CSN1	CSN1	343	332
5	CSN2	CSN2	150	157
6	CSN2	CSN2	110	150
7	CSN2	CSN2	253	215
8	CSN2	CSN2	415	364
9	CSN2	CSN2	426	364
10 [q]	CSN1	CSN1	467	467
11	CSN2	CSN2	243	157
12	CSN2	CSN2	56	43
13	CSN2	CSN2	64	93
14	CSN2	CSN2	81	48
15	CSN2	CSN2	48	64
16	CSN2	CSN2	48	93
17	CSN2	CSN2	93	73
18	CSN2	CSN2	64	73
19 [q]	CSN2	CSN2	415	415
20	CSN3	CSN3	254	243
21	CSN3	CSN3	125	115
22	CSN3	CSN3	254	281
23	CSN4	CSN4	227	214
24	CSN4	CSN4	214	188
25	CSN5	CSN5	308	179
26	CSN5	CSN5	193	218
27	CSN5	CSN5	218	294
28	CSN5	CSN5	218	298
29	CSN6	CSN6	75	113
30	CSN6	CSN6	229	236

q Denotes three intra-subunit cross-links to the same residue. Also omitted from Supplementary Figure 6.26b-c

31 [r]	CSN7b	CSN7b	221	217
32 [r]	CSN7b	CSN7b	221	257
33 [r]	CSN7b	CSN7b	217	199
34 [r]	CSN7b	CSN7b	218	217
35	CUL2	CUL2	393	382
36	CUL2	CUL2	179	252
37	CUL2	CUL2	393	404
38	CUL2	CUL2	677	720
39	CUL2	CUL2	720	382
40	CUL2	CUL2	728	677
41	CUL2	CUL2	252	244
42	CUL2	CUL2	433	386
43	CUL2	CUL2	608	615
44	CUL2	CUL2	238	244
45	CUL2	CUL2	404	400
46	CUL2	CUL2	433	677
47	CUL2	CUL2	62	149
48	CUL2	CUL2	179	96
49	CUL2	CUL2	433	472
50	CUL2	CUL2	677	382
51	ELOB	ELOB	19	36
52	ELOB	ELOB	36	28
53	ELOB	ELOB	36	46
54	ELOB	ELOB	11	28
55	ELOB	ELOB	11	36
56	ELOB	ELOB	46	28
57 [q]	ELOB	ELOB	104	104
58	NEDD8	NEDD8	60	6
59	NEDD8	NEDD8	11	33
60	VHL	VHL	196	171
Inter-subunit				
1	CSN1	CSN2	438	426
2	CSN1	CSN2	438	415
3 [r]	CSN1	CSN3	467	406
4	CSN2	CSN3	426	243

r Indicates six cross-links omitted from projection onto the CSN-CRL2~N8 structure in **Supplementary Figure 6.26** due to the lack of structural coverage

5	CSN2	CSN4	361	337
6	CSN2	CSN4	376	337
7	CSN2	CSN7b	426	199
8	CSN3	CSN8	296	58
9	CSN3	CSN8	115	65
10	CSN3	CSN8	115	58
11	CSN5	CSN7b	308	199
12	CSN5	CSN7b	179	199
13[r]	CSN5	CSN7b	294	217
14	CUL2	NEDD8	382	33
15	CUL2	NEDD8	433	6
16	CUL2	VHL	114	196
17	CUL2	ELOC	96	43
18	CUL2	ELOC	114	43
Inter-complex				
19	CSN1	ELOB	158	19
20	CSN2	CUL2	263	462
21	CSN2	CUL2	157	489
22	CSN2	CUL2	225	462
23	CSN2	CUL2	64	404
24	CSN4	RBX1	200	105

Supplementary Data 6.2. Chemical cross-links of the CSN-CRL2 complex.

Index	Protein A	Protein B	Residue A	Residue B
Intra-subunit				
1	CSN1	183	CSN1	153
2	CSN1	444	CSN1	451
3	CSN2	64	CSN2	93
4	CSN2	110	CSN2	150
5	CSN2	225	CSN2	189
6	CSN2	225	CSN2	253
7	CSN2	253	CSN2	215
8	CSN2	253	CSN2	300
9	CSN2	415	CSN2	364
10	CSN2	426	CSN2	415
11	CSN2	150	CSN2	157
12	CSN2	157	CSN2	243
13	CSN2	48	CSN2	81
14	CSN2	48	CSN2	93
15	CSN3	125	CSN3	115
16	CSN3	254	CSN3	243
17	CSN3	254	CSN3	281
18	CSN4	214	CSN4	200
19	CSN4	150	CSN4	200
20	CSN4	188	CSN4	214
21	CSN5	191	CSN5	219
22	CSN5	194	CSN5	219
23	CSN5	180	CSN5	309
24 [s]	CSN7b	221	CSN7b	217
25	CUL2	62	CUL2	149
26	CUL2	179	CUL2	96
27	CUL2	179	CUL2	252
28	CUL2	252	CUL2	244
29	CUL2	261	CUL2	244
30	CUL2	393	CUL2	404
31	CUL2	608	CUL2	615
32	CUL2	659	CUL2	404
33	CUL2	244	CUL2	238
34	CUL2	400	CUL2	404
35	CUL2	22	CUL2	74
36	CUL2	290	CUL2	291

s Indicates five cross-links omitted from projection onto the CSN-CRL2 structure in Supplementary Figure 6.30b-c due to lack of structural coverage

37	CUL2	244	CUL2	291
38	CUL2	400	CUL2	401
39	CUL2	677	CUL2	728
40	CUL2	719	CUL2	728
41[t]	CUL2	252	CUL2	252
42	ELOB	36	ELOB	46
43	ELOB	11	ELOB	28
44	ELOB	11	ELOB	36
45	ELOB	11	ELOB	46
46	ELOB	28	ELOB	36
47	VHL	196	VHL	171
Inter-subunit				
1	CSN1	422	CSN2	415
2	CSN1	438	CSN2	426
3	CSN1	438	CSN3	312
4[s]	CSN1	467	CSN7b	257
5	CSN2	303	CSN4	200
6	CSN2	426	CSN3	243
7	CSN2	361	CSN4	337
8	CSN3	115	CSN8	58
9	CSN3	115	CSN8	65
10	CSN4	372	CSN6	75
11	CSN5	180	CSN7b	199
12	CSN5	309	CSN7b	199
13[s]	CSN5	294	CSN7b	217
14	CUL2	114	VHL	196
15	CUL2	114	ELOC	43
16	ELOB	19	ELOC	32
17	ELOB	36	ELOC	20
18	ELOB	19	ELOC	6
Inter-complex				
1	CSN1	467	VHL	171
2	CSN2	225	CUL2	462
3	CSN2	263	CUL2	462
4	CSN2	415	CUL2	692
5[s]	CSN7b	218	CUL2	149
6[s]	CSN7b	221	CUL2	291

t Denotes one intra-subunit cross-links to the same residue. Also omitted from Supplementary Figure 6.30

Supplementary Data 6.3. Chemical cross-links of the CSN complex.

Index	Protein A	Protein B	Residue A	Residue B
Intra-subunit				
1	CSN1	136	CSN1	144
2	CSN1	144	CSN1	137
3	CSN1	183	CSN1	153
4	CSN1	224	CSN1	97
5	CSN1	422	CSN1	444
6	CSN1	444	CSN1	451
7 [u]	CSN1	467	CSN1	467
8	CSN2	48	CSN2	93
9	CSN2	64	CSN2	48
10	CSN2	64	CSN2	93
11	CSN2	64	CSN2	143
12	CSN2	93	CSN2	143
13	CSN2	93	CSN2	157
14	CSN2	110	CSN2	73
15	CSN2	110	CSN2	93
16	CSN2	110	CSN2	150
17	CSN2	150	CSN2	143
18	CSN2	150	CSN2	157
19	CSN2	157	CSN2	150
20	CSN2	157	CSN2	170
21	CSN2	167	CSN2	157
22	CSN2	170	CSN2	143
23	CSN2	225	CSN2	253
24	CSN2	253	CSN2	215
25	CSN2	415	CSN2	361
26	CSN2	415	CSN2	364
27	CSN2	426	CSN2	364
28	CSN2	426	CSN2	415
29 [u]	CSN3	115	CSN3	115
30	CSN3	125	CSN3	115
31	CSN3	254	CSN3	115
32	CSN3	254	CSN3	237
33	CSN3	254	CSN3	243
34	CSN3	254	CSN3	281
35	CSN3	281	CSN3	125
36	CSN3	281	CSN3	305

[u] Denotes intra-subunit cross-links to the same residue.

37 [v]	CSN4	32	CSN4	20
38 [v]	CSN4	32	CSN4	25
39 [v]	CSN4	116	CSN4	200
40 [v]	CSN4	137	CSN4	150
41 [v]	CSN4	137	CSN4	214
42 [v]	CSN4	144	CSN4	200
43 [v]	CSN4	150	CSN4	200
44 [u]	CSN4	200	CSN4	200
45 [v]	CSN4	214	CSN4	116
46 [v]	CSN4	214	CSN4	200
47 [v]	CSN4	251	CSN4	200
48 [v]	CSN4	290	CSN4	200
49	CSN5	179	CSN5	301
50	CSN5	229	CSN5	179
51	CSN5	301	CSN5	294
52	CSN6	75	CSN6	113
53	CSN6	117	CSN6	75
54	CSN7b	97	CSN7b	165
55	CSN7b	144	CSN7b	199
56	CSN7b	221	CSN7b	217
Inter-subunit				
1	CSN1	467	CSN3	115
2	CSN1	224	CSN5	294
3	CSN1	467	CSN7b	257
4	CSN2	415	CSN1	467
5	CSN2	426	CSN1	262
6	CSN2	426	CSN1	264
7	CSN2	426	CSN1	438
8	CSN2	415	CSN3	115
9	CSN2	426	CSN3	243
10	CSN2	441	CSN3	243
11 [v]	CSN2	253	CSN4	200
12 [v]	CSN2	300	CSN4	200
13 [v]	CSN2	361	CSN4	337
14 [v]	CSN2	415	CSN4	200
15	CSN2	415	CSN5	179
16	CSN3	243	CSN1	262
17	CSN3	281	CSN1	467
18	CSN3	312	CSN1	438

v Denotes CSN4 cross-links measured in **Supplementary Figure 6.31**

19	CSN3	254	CSN2	426
20	CSN3	391	CSN7b	199
21 [v]	CSN4	214	CSN2	364
22	CSN4	314	CSN6	75
23	CSN5	179	CSN7b	199
24	CSN6	75	CSN4	200
25	CSN6	243	CSN7b	164
26	CSN7b	217	CSN5	294
27	CSN8	166	CSN1	250
28	CSN8	58	CSN3	115
29	CSN8	65	CSN3	243
30	CSN8	65	CSN3	281